HOCHSCHULE
MITTWEIDA
UNIVERSITY OF
APPLIED SCIENCES

Masterarbeit

Herr B. Sc.
Eric Zuchantke

Datenvisualisierung in der CyanoFactory Knowledge Base

Mittweida, 2014

Fakultät Mathematik, Naturwissenschaften,
Informatik

Masterarbeit

Datenvisualisierung in der CyanoFactory Knowledge Base

Autor:
Herr B. Sc.

Eric Zuchantke

Studiengang:
Molekularbiologie/Bioinformatik

Seminargruppe:
MO12w1-M

Erstprüfer:
Prof. Dr. rer. nat. habil. Röbbe Wünschiers

Zweitprüfer:
M. Sc. Gabriel Kind

Einreichung:
Mittweida, 18.08.2014

Bibliographische Beschreibung

Zuchantke, Eric:

Datenvisualisierung in der CyanoFactory Knowledge Base. – 2014. – 7, 53 S.

Hochschule Mittweida, Fakultät Mathematik/Naturwissenschaften/Informatik

Masterarbeit, 2014

Englischer Titel:

Data visualization in the CyanoFactory Knowledge Base

Referat

Diese Masterarbeit befasst sich mit der Datenvisualisierung innerhalb der CyanoFactory Knowledge Base. Die CyanoFactory Knowledge Base ist dabei eine Webanwendung für die Speicherung, Verarbeitung und Visualisierung von Informationen bezüglich der Daten aus dem europäischen Forschungsprojekt CyanoFactory. Innerhalb der Arbeit wurden dabei die interaktiven Visualisierungen von Protein-Protein-Interaktionen und Protein-Chemikalien-Interaktionen des Modellorganismus *Synechocystis sp. PCC 6803* umgesetzt und als Webtool in die CyanoFactory Knowledge Base integriert. Im Rahmen dieser Arbeit wurde auch eine visuelle Ausgabe für Flux-Balance-Analysen erstellt. Neben der Visualisierung der Interaktionen innerhalb von *Synechocystis sp. PCC 6803* wurde eine Analyse von dessen Protein-Interaktionsnetzwerk durchgeführt.

Abstract

This master thesis is about the data visualization within the CyanoFactory Knowledge Base. The CyanoFactory Knowledge Base is a web application for storing, processing and visualization of information relating to the data from the European research project CyanoFactory. Within this work, an interactive visualization of protein-protein interactions and protein-chemical interactions of the model organism *Synechocystis sp. PCC 6803* are implemented and integrated into the web tool CyanoFactory Knowledge Base. Furthermore, in this work a visual output for Flux-balance analysis was created. In addition to the visualization of interactions within *Synechocystis sp. PCC 6803*, an analysis of the interaction-network of *Synechocystis sp. PCC 6803* was performed.

Danksagung

Ich möchte mich zu aller erst bei Herrn Prof. Dr. rer. nat. habil. Röbbe Wünschiers für die Möglichkeit des Mitwirkens an dem CyanoFactory Projekt bedanken und seine Unterstützung während der gesamten Zeit. Des Weiteren gilt mein Dank auch meinem Betreuer Herrn M.Sc. Gabriel Kind für seine Hilfe und Ratschläge bei den kleineren und größeren informatischen Problemen.

Allgemein möchte ich mich bei der gesamten Fachgruppe Biotechnologie für deren freundliche Unterstützung während der gesamten Zeit bedanken.

Ein weiterer großer Dank gilt meiner Familie und Freunden welche mich über das gesamte Studium begleitet haben.

Natürlich möchte ich mich auch bei Susanne Bause für ihren Rat und ihre Geduld bei der Korrekturlesung dieser Arbeit bedanken.

Inhaltsverzeichnis

Abbildungsverzeichnis

Tabellenverzeichnis

Abkürzungsverzeichnis

Abkürzung	Bedeutung
D3	Data-Driven Documents
FBA	Flux Balance Analyse
GTP	Guanosintriphophat
KB	Knowledge Base
PDF	Portable Document Format
PNG	Portable Network Graphics
PyNetMet	Python Network Metabolism
STITCH	Search Tool for Interacting Chemicals
STRING	Search Tool for the Retrieval of Interacting Genes/Proteins
SVG	Scalable Vector Grafik

1 Einleitung

„There is a magic in graphs. The profile of a curve reveals in a flash a whole Situation - the life history of an epidemic, a panic, or an era of prosperity. The curve informs the mind, awakens the imagination, convinces.

Graphs carry the message home. A universal language, graphs convey information directly to the mind. Without complexity there is imaged to the eye a magnitude to be remembered. Words have wings, but graphs interpret. Graphs are pure quantity, stripped of verbal sham, reduced to dimension, vivid, unescapable.

Graphs are all inclusive. No fact is too slight or too great to plot to a scale suited to the eye. Graphs may record the path of an ion or the orbit of the sun, the rise of a civilization, or the acceleration of a bullet, the climate of a century or the varying pressure of a child.

The graphic art depicts magnitudes to the eye. It does more. It compels the seeing of relations. We may portray by simple graphic methods whole masses of intricate routine, the organization of an enterprise, or the plan of a campaign. Graphs serve as storm signals for the manager, statesman, engineer; as potent narratives for the actuary, statist, naturalist; and as forceful engines of research for science, technology and industry. They display results. They disclose new facts and laws. They reveal discoveries as the bud unfolds the flower,

The graphic language is modern. We are learning its alphabet. That it will develop a lexicon and a literature marvelous for its vividness and the variety of application is inevitable.

Graphs are dynamic, dramatic. They may epitomize an epoch each dot a fact, each slope an event, each curve a history. Wherever there are data to record, inferences to draw, or facts to tell, graphs furnish the unrivalled means whose power we are just beginning to realize and to apply.”

Mit diesen Worten beschreibt Henry D. Hubbard im Vorwort des Buches „Graphic Presentation“ [Brinton, 1939], welcher Nutzen und welche Schönheit in Graphen jeglicher Form stecken. Dabei handelt es sich bei Henry D. Hubbard um keinen Informatiker oder Naturwissenschaftler, sondern um einen Juristen und Politiker.

Die visuelle Darstellung von Daten ermöglicht es, diese einem breiten Spektrum verständlich zu machen. Somit können auch neue Ideen oder auch andere Herangehensweisen in ein Projekt einfließen.

In den letzten Jahren ist die Menge an produzierten Daten drastisch angestiegen. Es wird deshalb oftmals in diesem Zusammenhang die Beschreibung „Big Data“ genutzt.

Der Bereich in dem diese Menge an Daten erzeugt wird, ist dabei breit gefächert. Neben Unternehmen wie Facebook, Google und Apple, wird auch in wissenschaftlichen Bereichen eine hohe Datenlast erzeugt. Dabei kann durch verschiedene Experimente eine Unmenge an Daten anfallen. Bei den anfallenden Daten ist jedoch nicht sicher in wie weit diese relevant sind und es muss deshalb schon bei der Speicherung der Daten eine Filterung erfolgen. Wie stark der Informationsgehalt der Daten ist, hängt natürlich vom Betrachter und der Fragestellung bzw. dem Ziel einer jeden Untersuchung ab.

Diese Arbeit soll in gewissem Sinne an die Idee von Verständlichkeit für Jedermann anknüpfen. Es sollen die gewonnenen Daten aus dem europäischen Projekt „CyanoFactory“ aufgearbeitet und der Informationsgehalt der Daten visualisiert werden. Hierbei soll mit verschiedenen mathematischen, informatischen Ansätzen gearbeitet werden, um dieses Ziel zu erreichen. Im Endeffekt sollen somit alle Partner des „CyanoFactory“-Projektes in der Lage sein, sämtliche Ergebnisse der anderen Partner nutzen zu können und, wie es schon Henry D. Hubbard beschrieb, die Schönheit der Daten zu sehen.

Aufgrund der vorliegenden Daten zum Zeitpunkt dieser Arbeit, beschränkt sich die Visualisierung auf die Darstellung von Interaktionen zwischen Proteinen und Chemikalien. Des Weiteren soll die Modellierung von metabolischen Prozessen visualisiert werden.

1.1 CyanoFactory

Seit Jahren ist man auf der Suche nach neuen Kraftstoffquellen, um die begrenzten fossilen Brennstoffe zu ersetzen. Das seit 2013 geförderte EU-Projekt „CyanoFactory“ ist dabei an der Suche beteiligt und versucht auf biotechnologische Art und Weise Wasserstoff mithilfe von genetisch modifizierten Cyanobakterien, *Synechocystis sp. PCC6803,* kosteneffizient zu erzeugen. CyanoFactory ist dabei eine Zusammenarbeit aus zehn unabhängigen Forschungspartnern aus verschiedenen europäischen Ländern. Die beteiligten Forschungspartner sind mit ihren Aufgaben und Standorten in Tabelle 1 beschrieben. [Kind, 2013] [URL-1]

Tabelle 1 Übersicht der CyanoFactory Projektpartner mit dem jeweiligen Standort und Aufgabe. Unterstrichen ist dabei der Standort des Leiters nach [URL-1].

Partner	Land	Aufgabe
Uppsala Universitet	Schweden	Toolbox für synthetische Biologie bei Cyanobakterien
Hochschule Mittweida	Deutschland	Data Warehouse
Ruhr-Universität Bochum	Deutschland	Verbesserte Photosynthese-Effizienz für H2-Produktion
KSD Innovation GmbH	Deutschland	Entwicklung eines effizienten Photobioreaktors
CNR-ISE	Italien	Prototyp-Photobioreaktor,
M2M Engineering S.A.S.	Italien	Herstellung und Leistungstest
Instituto de Biologia Molecular e Celular	Portugal	Verbesserung Chassis-Wachstum Funktionalität und Zuverlässigkeit
Universa v Ljubljani	Slowenien	Biologische Sicherheit
Universidad Politécnica de Valencia	Spanien	Metabolische Modellierung der konstruierten Zellen
University of Sheffield	Vereinigtes Königreich	Untersuchung konstruierter Zellen auf Engpässe

1.2 Synechocystis sp. PCC6803

Das Cyanobakterium *Synechocystis sp. PCC6803* wurde 1968 in die Pasteur Culture Collection eingepflegt und wird weltweit als Modellorganismus genutzt. *Synechocystis sp. PCC6803* war der erste photosynthetische Organismus, dessen Genom vollständig sequenziert wurde. Das Chromosom enthält dabei 3264 mögliche protein-codierende Gene. Neben dem Chromosom wurden auch noch vier unterschiedlich große Plasmide vollständig sequenziert. Dieses Cyanobakterium besitzt dabei die spezielle Eigenschaft spontan transformierbar zu sein und ist somit fähig, fremde DNA in sein Genom zu integrieren. Dies ermöglicht ein Überleben bei unterschiedlichsten Lebenseinflüssen. [Kaneko et al., 2003; Kind, 2013]

1.3 CyanoFactory Knowledge Base

Da es sich bei dem Projekt CyanoFactory um ein europaweites Projekt handelt, ist es notwendig alle gewonnenen und benötigten Daten allen Projektpartner ohne längere Wartezeiten zur Verfügung zu stellen. Deshalb ist es wichtig, dass die Daten nicht nur zentral und für alle zugänglich gelagert werden. Somit ist ein schneller Informationsaustausch bzw. Informationsgewinn zwischen den einzelnen Projektpartnern gewährleistet. Hierfür wird ein sogenanntes Data Warehouse genutzt, die CyanoFactory Knowledge Base [Kind, 2013]. Die

Aufgabe dieser Knowledge Base (KB) ist dabei jedoch nicht nur das reine Sammeln von Informationen, sondern auch die Aufarbeitung der Informationen aus den einzelnen Quellen. Dabei werden die Informationen in ein einheitliches Format überführt und mittels einer relationalen Datenbank, PostgreSQL, gespeichert. Mittels eines Webinterfaces wird anschließend den einzelnen Projektpartner der Zugang zu der CyanoFactory KB ermöglicht. Durch dieses Webinterface werden die verarbeiteten Daten visuell dargestellt werden. Die Funktionsweise wird schematisch in Abbildung 1 dargestellt.

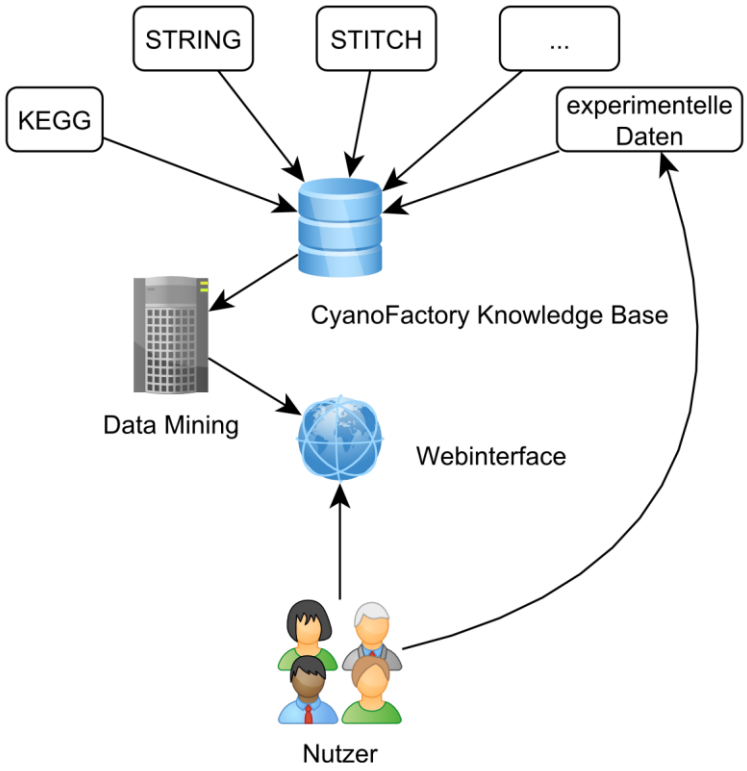


Abbildung 1 Funktionsweise CyanoFactory Knowledge Base. Durch die Nutzung verschiedener Quellen, wie die Datenbanken von KEGG, STRING, STITCH und weiteren, sowie experimentellen Daten der Projektpartner wird die Datenbank der CyanoFactory KB gespeist. Die Informationen werden analysiert und verarbeitet und durch ein Webinterface den Projektpartnern zur Verfügung gestellt.

Derzeit stehen die gesammelten Information zu sämtlichen bekannten Genen und deren Produkten, sowie die Lage dieser auf dem Chromosom und den Plasmiden zur Verfügung. Des Weiteren können die metabolischen Wege verschiedener Stoffe innerhalb von *Synechocystis sp. PCC6803* betrachtet werden.

Die CyanoFactory Knowledge Base basiert in ihrem Aufbau, auf der WholeCell Knowledge Base [Karr et al., 2013; Kind, 2013].

1.4 STRING

Das Search Tool for the Retrieval of Interacting Genes/Proteins (STRING) ist sowohl eine webbasierte Anwendung, als auch eine Datenbank für die Interaktionen von Proteinen bzw. Genen verschiedener Organismen [Mering et al., 2005]. Dabei trägt STRING die Informationen diverser Quellen zusammen und bewertet diese. In der Version 9.1 stellt die STRING Datenbank Interaktionsinformationen für mehr als 5 Proteine von über 1000 Organismen zur Verfügung [Franceschini et al., 2013].

Die Bewertung einer Interaktion wird anhand eines Interaktionswertes durchgeführt. Dieser ergibt sich aus den in Tabelle 2 beschriebenen einzelnen Nachweisen: Datenbanken, Experimenten, Gen-Fusionen, Kookkurrenzen, Koexpressionen, Nachbarschaften und Textmining. Der Interaktionswert ist dabei lediglich eine Angabe, wie wahrscheinlich es ist, dass das Protein A mit Protein B interagiert. Neben den existierenden Informationen zu Interaktionenen zwischen Proteinen, werden auch Vorhersagen zu Protein-Protein-Interaktionen durchgeführt. Diese Vorhersagen basieren auf den vorliegenden Informationen zu den jeweiligen Proteinen, bzw. auf Homologien zu anderen Organismen.

Tabelle 2 Interaktionstypen der STRING Datenbank mit kurzer Erläuterung des jeweiligen Types nach [Franceschini et al., 2013; Mering et al., 2005; Szklarczyk et al., 2010; Jensen et al., 2009]

Interaktionstyp	Beschreibung
Datenbanken	Überprüfung von anderen Datenbanken
Experimente	Experimentelle Versuche zur Protein-Protein-Interaktion
Gen-Fusion	Zusammenschluss von mehreren Genen eines Genoms zu einem einzigen Gen in einem anderen Genom.
Homologie	Interaktionen von Proteinen welche in anderen Organismen vorkommen und auf den ausgewählten Organismus übertragen werden können
Kookkurrenz	Auftreten von Genen im gleichen Stoffwechsel oder Organismus
Koexpression	Transkriptionsverhältnis ähnelt sich bei Genen unter verschiedenen Umständen
Nachbarschaft	Betrachtung der physikalischen Entfernung von Genen auf dem Chromosom. Je dichter Gene aneinander liegen, desto wahrscheinlicher ist eine funktionelle Beziehung
Textmining	Automatische Untersuchung von wissenschaftlichen Texten auf ein gemeinsames Auftreten verschiedener Gene bzw. Proteine.

Die Interaktionswerte werden als ganze Zahlen abgelegt. Die Nutzung von ganzen Zahlen innerhalb der Daten erfüllt dabei den Zweck bei der Berechnung des Interaktionswertes Gleitkommazahlen zu vermeiden, um damit Rundungsfehlern vorzubeugen. Die Berechnung des Interaktionswertes S erfolgt auf einer naiven Bayes Weise, dargestellt in (1.1). Dabei sind unter S_i die Score-Werte der einzelnen Nachweise beschrieben [Mering et al., 2005].

$$S = 1 - \prod_i (1 - S_i) \tag{1.1}$$

Für die Berechnung wird dabei der vorhandene Score-Wert in einen Wert zwischen null und eins überführt, durch die Division mit 1000. Nach der Berechnung wird der Interaktionswert wiederum mit 1000 multipliziert. Somit gilt $0 < S < 1000$.

Neben der Datenbank besitzt STRING, wie anfangs erwähnt, ebenfalls eine webbasierte Anwendung, welche das Suchen nach einem Protein und dessen Interaktionen ermöglicht. Diese Anwendung gibt dabei einen Graph, wie in Abbildung 2 dargestellt ist, zurück. Dieser zeigt durch die Kantenfarben die jeweiligen Quellen der Interaktionen an.

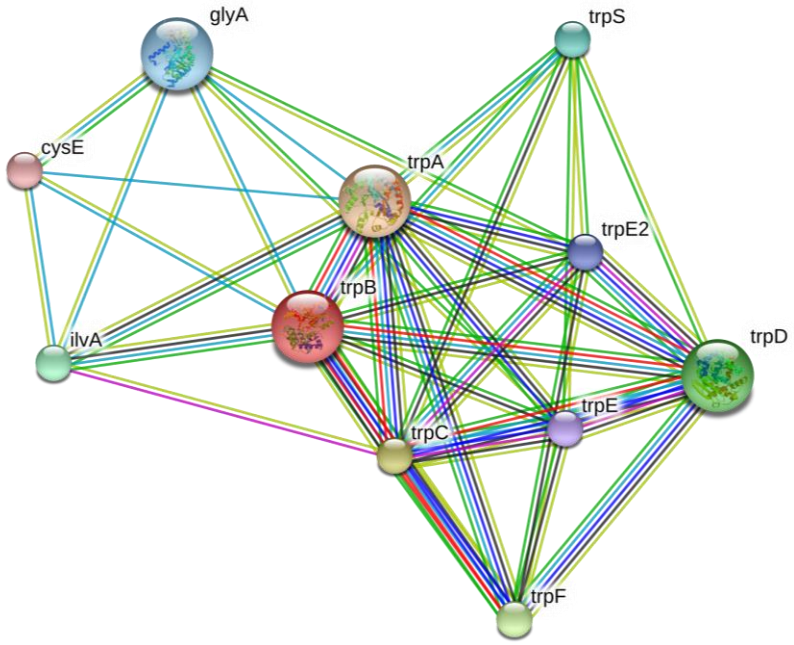
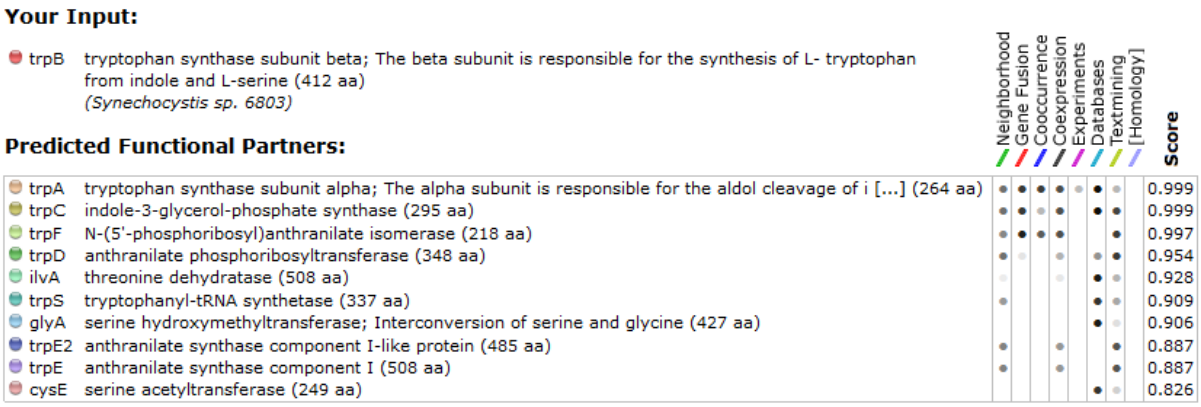


Abbildung 2 Darstellung von Protein-Protein-Interaktionen über STRING. Als Ausgangsprotein wurde trpB gewählt. Jede Kante stellt einen spezifischen Interaktionstyp dar [URL-2].

Neben dem Graphen gibt es auch eine tabellarische Ausgabe, siehe Abbildung 4, welche nochmals die interagierenden Proteine und die Quellen der Interaktionen beschreibt. STRING bietet die Möglichkeit eigene Interaktionsdaten hochzuladen und somit das gesamte Interaktionsnetzwerk zu erweitern.



Your Input:

		Neighborhood	Gene Fusion	Cooccurrence	Coexpression	Experiments	Databases	Textmining	[Homology]	Score
trpB	tryptophan synthase subunit beta; The beta subunit is responsible for the synthesis of L- tryptophan from indole and L-serine (412 aa) *(Synechocystis sp. 6803)*									

Predicted Functional Partners:

trpA	tryptophan synthase subunit alpha; The alpha subunit is responsible for the aldol cleavage of i [...] (264 aa)	●	●	●	●	●	●	●		0.999
trpC	indole-3-glycerol-phosphate synthase (295 aa)	●	●	●	●		●	●		0.999
trpF	N-(5'-phosphoribosyl)anthranilate isomerase (218 aa)	●	●	●	●			●		0.997
trpD	anthranilate phosphoribosyltransferase (348 aa)	●	●		●		●	●		0.954
ilvA	threonine dehydratase (508 aa)						●	●		0.928
trpS	tryptophanyl-tRNA synthetase (337 aa)	●					●	●		0.909
glyA	serine hydroxymethyltransferase; Interconversion of serine and glycine (427 aa)						●	●		0.906
trpE2	anthranilate synthase component I-like protein (485 aa)	●			●			●		0.887
trpE	anthranilate synthase component I (508 aa)	●			●			●		0.887
cysE	serine acetyltransferase (249 aa)						●	●		0.826

Abbildung 3 Informationen zu Interaktionspartner von trpB. Es werden Informationen zu den Interaktionspartnern gezeigt, wie deren Namen, eine kurze Beschreibung, die Anzahl an Aminosäuren und die Interaktionstypen, sowie den Interaktionswert (Score) [URL-2].

1.5 STITCH

Die Datenbank und Webanwendung Search Tool for Interacting Chemicals (STITCH) bietet, wie die in Abschnitt 1.3 beschriebene STRING, ähnliche Möglichkeiten zur Analyse bzw. Angabe von Interaktionen [Kuhn et al., 2014]. Jedoch erweitert STITCH die Möglichkeiten von STRING, da es Chemikalien in die Suche nach Interaktionen mit einbezieht bzw. liegt das Hauptaugenmerk dabei auf den Interaktionen von Proteinen mit Chemikalien. Dabei werden auch Interaktionen zwischen Chemikalien und Proteinen vorhergesagt. Die für die STITCH Datenbank genutzten Quellen liegen wiederum bei den verschiedenen anderen Datenbanken. In der STITCH Version 4.0 stehen Informationen zu Interaktionen zwischen 2,6 Millionen Proteinen von 1133 Organismen und 300000 chemischen Verbindungen zur Verfügung. Die Interaktionen zwischen Proteinen und Chemikalien werden in Interaktionstypen eingeteilt. Diese sind: Experiment, Datenbank, Textmining und Vorhersage. Sie ergänzen die Protein-Protein-Interaktionstypen aus der STRING Datenbank, siehe Tabelle 2. Die Berechnung des Interaktionswertes erfolgt dabei wieder mittels der einzelnen Interaktionstypen wie in (1.1) beschrieben.

Grundsätzlich ist die Webapplikation von STITCH so ausgelegt, dass die Daten der externen Datenbanken zusammengefasst werden. Dies bietet die Möglichkeit Informationen zeiteffizient zu sammeln und auszuwerten. Dabei ist es allerdings jederzeit möglich zu den einzelnen Quellen der Informationen zurück zu gelangen. Des Weiteren werden die Interaktionen, wie bei STRING, mittels eines Graphen visuell dargestellt, wie in Abbildung 4 gezeigt ist.

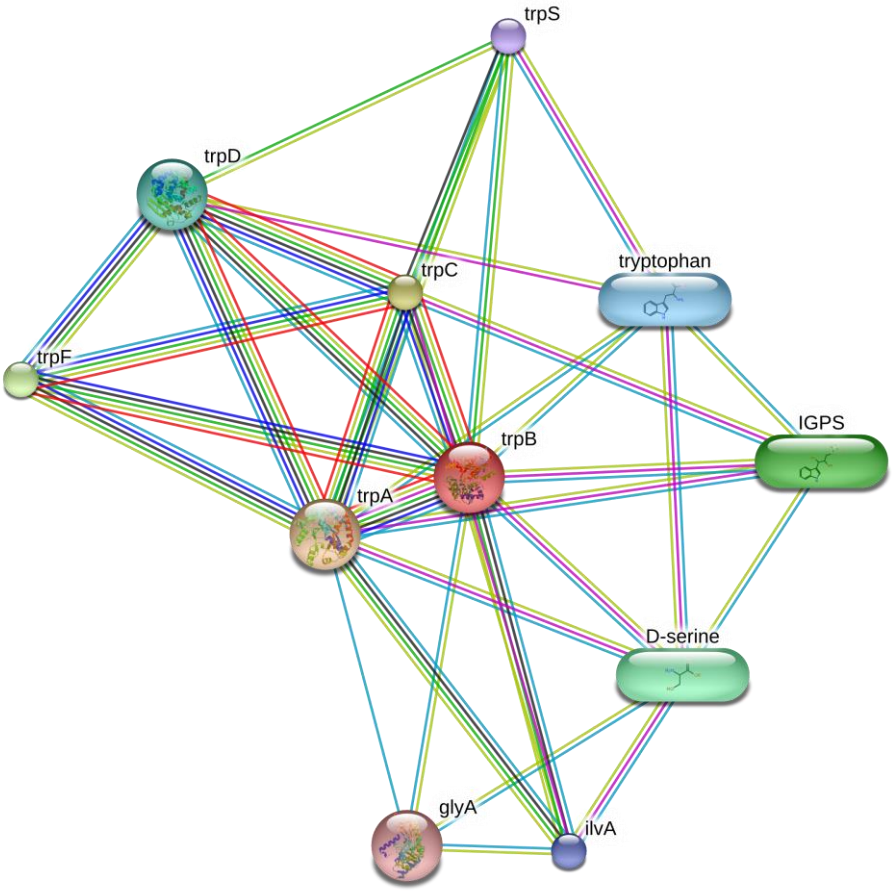


Abbildung 4 Interaktionen von trpB mit Proteinen und Chemikalien [URL-3]. Dabei sind die Chemikalien elliptisch und die Proteine kreisförmig dargestellt. Die einzelnen Linien zwischen den Proteinen und Chemikalien stellen deren Interaktionstypen dar.

Es werden die Interaktion ebenfalls in tabellarischer Form, Abbildung 5, aufgelistet. Ebenfalls wird die Möglichkeit geboten eigene Nutzerdaten hochzuladen und somit in das Interaktionsnetzwerk zu integrieren.

Your Input:

		Neighborhood	Gene Fusion	Cooccurrence	Coexpression	Experiments	Databases	Textmining	[Homology]	Score
trpB	tryptophan synthase subunit beta; The beta subunit is responsible for the synthesis of L- tryptophan from indole and L-serine (412 aa) *(Synechocystis sp. 6803)*									

Predicted Functional Partners:

trpA	tryptophan synthase subunit alpha; The alpha subunit is responsible for the aldol cleavage of i [...] (264 aa)	●	●	●	●	●	●	●		0.999
trpC	indole-3-glycerol-phosphate synthase (295 aa)	●	●	●	●		●	●		0.999
trpF	N-(5'-phosphoribosyl)anthranilate isomerase (218 aa)	●	●	●	●			●		0.997
trpD	anthranilate phosphoribosyltransferase (348 aa)	●	●		●		●	●		0.954
ilvA	threonine dehydratase (508 aa)	●			●		●	●		0.928
trpS	tryptophanyl-tRNA synthetase (337 aa)	●					●	●		0.909
glyA	serine hydroxymethyltransferase; Interconversion of serine and glycine (427 aa)						●	●		0.906
trpE2	anthranilate synthase component I-like protein (485 aa)	●			●			●		0.887
trpE	anthranilate synthase component I (508 aa)	●			●			●		0.887
cysE	serine acetyltransferase (249 aa)						●	●		0.826

Abbildung 5 Informationen zu trpB aus der STITCH [URL-3]. Wie bei STRING werden zusätzliche Informationen zu den Interaktionspartnern dargestellt. Diese beinhalten den Namen, eine kurze Beschreibung, die Sequenzlänge sowie die einzelnen Interaktionstypen und den Interaktionswert (Score).

1.6 Graphentheorie

Die Graphentheorie ist eines der wichtigsten Teilgebiete der diskreten Mathematik. Mit Hilfe der Graphentheorie lassen sich Probleme in verschiedenen netzartigen Strukturen beschreiben. Zu solchen Strukturen zählen Straßennetze, Computernetzwerke oder auch die in Abschnitt 1.3 und 1.5 beschriebenen Interaktionsnetzwerke. Die Art und Weise wie ein solches Netzwerk, auch Graph genannt, untersucht wird hängt von der gegebenen Problemstellung ab. In einem Graphen werden dabei die Objekte, wie Proteine, als Knoten aufgefasst und die Interaktionen zwischen den einzelnen Proteinen werden durch Kanten beschrieben.

Wir können Graphen dabei in zwei Kategorien unterscheiden. Zum einen den ungerichteten Graph G und den gerichteten Graph H. Beide Graphen bestehen aus der Knotenmenge V, unterscheiden sich jedoch in den Kantenmengen E und F voneinander. Somit ist $G = (V, E)$ und $H = (V, F)$. Der Unterschied zwischen diesen beiden Formen liegt in der Interpretation der Kanten. Bei dem ungerichteten Graph G werden die Knoten $a, b \in V$ durch die ungerichtete Kante $e = \{a, b\}$ verbunden. Dabei ist die Richtung der Kante nicht festgelegt. Es existiert somit keine Vorschrift wie die einzelnen Knoten abgelaufen werden sollen. Der gerichtete Graph H hingegen besitzt gerichtete Kanten. So beschreibt die Kante $f = (a, b)$ den gerichteten Verlauf vom Knoten a zum Knoten b. Somit kann man b nur von a aus erreichen, allerdings nicht a von b aus. [Tittmann, 2003]

Die graphische Darstellung erfolgt dabei für die Knoten meist als Kreise oder Punkte. Die Kanten zwischen den Knoten werden als Strecken oder Kurven zwischen den jeweiligen Kreisen, wie in Abbildung 6, dargestellt.

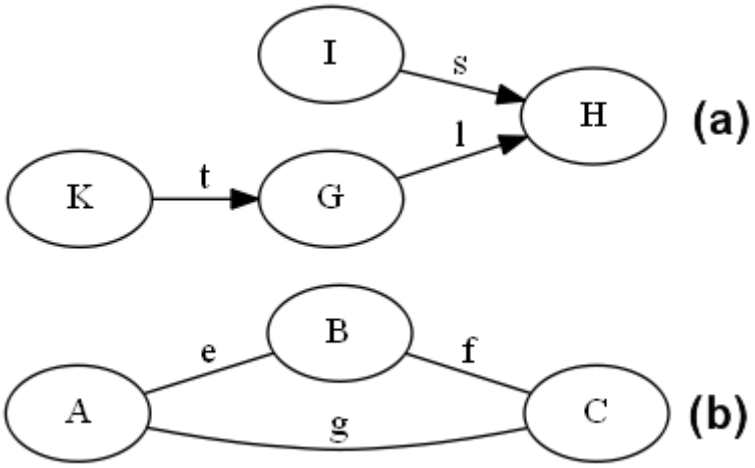


Abbildung 6 Darstellung von Graphen. (a) stellt einen gerichteten Graphen mit der Knotenmenge {K, G, I, H} und der Kantenmenge {s ,t ,l} dar. b zeigt einen ungerichteten Graphen mit der Kantenmenge {e, g, f} und der Knotenmenge {A, B, C}.

1.7 Flux Balance Analyse

Flux Balance Analysen (FBA) werden genutzt um die metabolischen Netzwerke einzelner Organismen zu rekonstruieren. Dabei beinhalten diese metabolischen Netzwerke alle Informationen, die zu den jeweiligen metabolischen Reaktionen bekannt sind. Durch solche Berechnungen lassen sich Vorhersagen bezüglich der Produktion eines Metaboliten oder des Wachstum eines Organismus, ausgehend von definierten Ausgangsbedingungen treffen [Orth et al., 2010].

Grundsätzlich werden dabei alle Reaktionen innerhalb einer stöchiometrischen Matrix S erfasst. S besitzt dabei die Größe $m \times n$. Jede Zeile $1, .., m$ in S stellt dabei einmalig einen gegebenen Metaboliten dar. Die Spalten $1, \dots, n$ stehen für die individuellen Reaktionen. Die einzelnen Werte $S_{m,n}$ der Matrix S stehen für die Koeffizienten des Metaboliten m in der Reaktion n. Ein positiver Koeffizient stellt dabei die Produktion des Metaboliten dar und ein negativer Koeffizient dagegen den Verbrauch eines Metaboliten. Da nicht alle Metabolite in einer einzigen enzymatischen Reaktion genutzt werden, werden deren Koeffizienten für diese Reaktion auf null gesetzt. Die Konzentrationen der Gesamtheit an Metaboliten werden durch den Vektor $\vec{x}$ beschrieben und dieser besitzt dabei die Länge m. Der Flux dieser Reaktionen wird durch den Vektor $\vec{v}$ dargestellt. Dieser besitzt die Größe n. Für die Berechnung der maximalen Biomasseproduktion, werden diese Vektoren für ein lineares Lösungssystem genutzt bzw. wird eine lineare Optimierung durchgeführt. Dabei geht man meist davon aus, dass sich das System im sogenannten Fließgleichgewicht befindet [Orth et al., 2010].

1.8 Python

„Beautiful is better than ugly. Explicit is better than implicit. Simple is better than complex. Complex is better than complicated. Flat is better than nested. Sparse is better than dense. Readability counts.“ Mit diesen Worten beginnt „The Zen of Python“ des Programmierers Tim Peters. Dieses stellt seiner Meinung nach das Regelwerk und somit die Richtlinien der Programmiersprache dar. Python gehört zu den höheren Programmiersprachen und ist multiparadigmatisch, welches bedeutet, dass Python mehrere Programmierparadigmen unterstützt. Diese sind Objektorientiertheit, Funktionalität und Prozeduralität. Dabei ist es mit Python möglich eigenständige Programme zu schreiben oder auch Skripte innerhalb verschiedener Anwendungsbereiche zu implementieren. Python gehört dabei mit zu den meist verbreiteten Programmiersprachen weltweit [Downey, 2012].

1.8.1 Django

Für die Erstellung dynamischer Webseiten bietet sich die Nutzung von Web Application Frameworks an. Diese Frameworks dienen unter anderem zur Vereinfachung sich wiederholender Tätigkeiten. Ein solches Framework ist das quelloffene Django.

Bei Django handelt es sich um ein Datenbank-basiertes Webframework, welches Python als Programmiersprache nutzt. Es wurde für die Verwendung in Nachrichtenredaktionen konzipiert, weshalb die Erstellung von Webinhalten einfach und schnell sein musste. Dadurch das Django quelloffen ist, wurde es auch für viele Unternehmen attraktiv. Zu diesen Unternehmen zählt beispielsweise die Washington Post und Google [Forcier, Bissex & Chun, 2008].

1.8.2 NetworkX

Python besitzt für die Erstellung, Bearbeitung und Analyse von Netzwerken und Graphen das Paket NetworkX [Hagberg et al., 2008]. Es ermöglicht eine schnelle und einfache Bearbeitung von Netzwerkinformation und stellt verschiedene Analysemethoden zur Verfügung. NetworkX ermöglicht die Darstellung von gerichteten und ungerichteten Graphen sowie Multigraphen und MultiDigraphen. Zusätzlich können Wichtungen an den Kanten angebracht werden sowie weitere optionale Eigenschaften für Knoten und Kanten.

1.8.3 PyNetMet

Wie NetworkX ist auch PyNetMet [Gamermann et al., 2014] ein Pythonpaket, welches für Flux Balance Analyse genutzt werden kann. PyNetMet steht dabei für Python Network Metabolism. Es bietet die Möglichkeit sowohl mit dem Dateiformat Optgene [Cvijovic et al., 2010] als auch mit dem Format SBML [Hucka et al., 2003] zu arbeiten. Dabei besteht PyNetMet aus den Grundeinheiten: Enzym, Network, Metabolism und FBA. Neben der FBA verfügt PyNetMet

dabei noch über die Möglichkeit, metabolische Netzwerke mittels graphentheorischer Ansätze zu analysieren.

1.9 JavaScript

Bei JavaScript handelt es sich um eine Skriptsprache. Skriptsprachen wie JavaScript unterscheiden sich dabei von Compilersprachen wie Java. Grundsätzlich werden die einzelnen Sprachelemente von JavaScript mittels des Interpreters verarbeitet, jedoch nicht in Maschinensprache übersetzt [Hirsemann & Rochusch, 2003]. Im Falle von JavaScript ist der Interpreter ein Browser wie beispielsweise Firefox, Safari oder Chrome.

Durch das Interpretieren der JavaScripts sind diese im Vergleich zu kompilierten Java-Anwendungen in Maschinensprache langsamer. Jedoch ist dieser Geschwindigkeits-unterschied durch die verbesserte technische Ausstattung von Computern nicht mehr relevant.

Ein großer Vorteil von JavaScript ist die Möglichkeit dynamische HTML-Seiten zu generieren und diese auch interaktiv zu gestalten. Dabei funktioniert JavaScript plattformübergreifend, jedoch existieren teilweise Unterschiede bei der Interpretation der verschiedenen Browser [Hirsemann & Rochusch, 2003].

1.9.1 D3

D3 steht für Data-Driven Documents und ist eine JavaScript Bibliothek, mit deren Hilfe man dokumentbasierte Daten manipulieren kann [Michael Bostock et al., 2011]. Hiermit ist es möglich die Elemente einer Webseite interaktiv zu gestalten. Damit sind sowohl die reinen HTML- und CSS-Teile einer Seite als auch eingebundene Objekte, wie beispielsweise Scalable Vector Grafiken (SVG), gemeint. Dabei bietet diese Bibliothek eine große Anzahl an Möglichkeiten Daten zu visualisieren. So können auf einfachste Weise Graphen bzw. Netzwerke mit physikalischen Modellen generiert werden. Weiterhin können auch simple Diagramme geplottet bzw. animiert werden und es stehen viele weitere Möglichkeiten zur Verfügung. Eine Auswahl an Visualisierungsmöglichkeiten ist in Abbildung 7 gegeben.

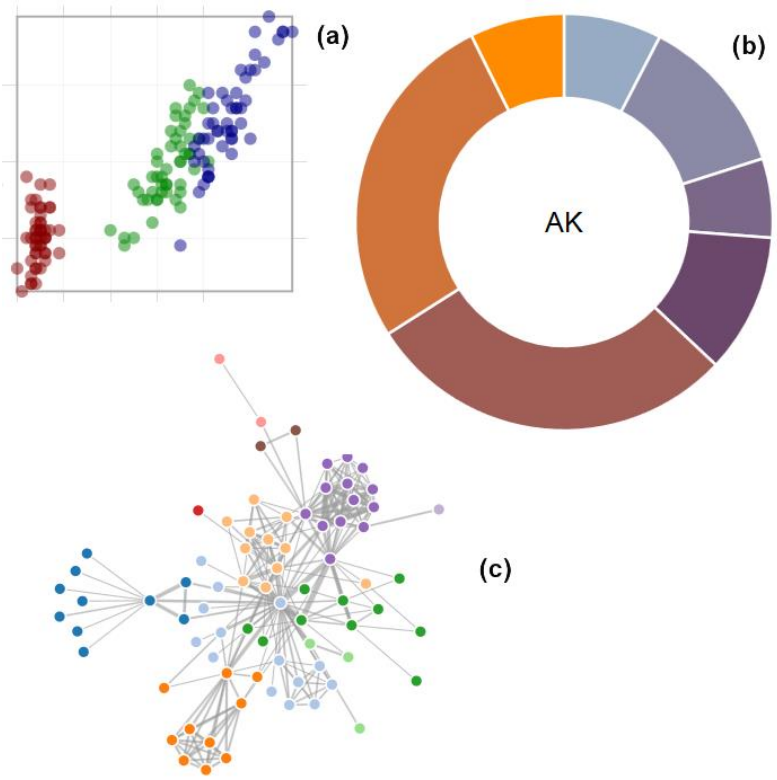


Abbildung 7 Drei Darstellungsmöglichkeiten von D3, hierbei ist nur eine kleine Auswahl an Möglichkeiten getroffen. (a) ist eine Punktwolke, (b) ein Kreisdiagramm und (c) ein Kraft-Layout Graph [URL-4].

Der große Vorteil von D3 und damit auch JavaScript ist die Interaktivität. Durch diese werden keine starren Abbildungen generiert, sondern es besteht die Möglichkeit Diagramme nachträglich zu bearbeiten bzw. den angezeigten Bereich zu ändern. Durch diese dynamischen Darstellungsmöglichkeiten ist es möglich die Informationsdichte um ein Vielfaches zu steigern beziehungsweise können Informationen dadurch vielschichtiger dargestellt werden.

1.10 GraphViz

GraphViz ist ein Visualisierungstool, welches von den Bell-Labs und AT&T entwickelt wurde [Gansner & North, 2000]. Dieses steht als Open Source einem breiten Spektrum von Nutzern zur Verfügung. Es ermöglicht sowohl einfache kleinere Darstellungen von Fließdiagrammen, als auch komplexen Netzwerken. Als ein komplexes Netzwerk können beispielsweise die Prozessabläufe innerhalb einer Zelle gesehen werden.

Dabei bietet GraphViz die Möglichkeit, die generierten Graphen, Netzwerke und Fließdiagramme in mehreren Formaten zu speichern. Zu diesen Formaten zählen unter anderem die SVG, das Portable Document Format (PDF), die Portable Network Graphics (PNG) und noch viele weitere.

Der Aufbau einer Datei, welche für das Erzeugen eines Graphen erforderlich ist, ist dabei einfach gehalten. Grundsätzlich gibt es bei GraphViz die Möglichkeit vier unterschiedliche Arten von Graphen zu generieren. Dazu zählen die Formen „graph“ und „digraph“ sowie die Formen „strict digraph“ und „strict graph“. Durch die Nutzung von „strict“ wird ein Graph so definiert, dass jeweils nur eine Kante zwei Knoten verbinden darf. Ohne die Präambel „strict“ sind Mehrfach-Kanten möglich. Unter der Form „digraph“ wird ein gerichteter Graph und mittels „graph“ ein ungerichteter Graph generiert. Bei dem gerichteten Graph besitzt jede Kante einen Pfeilkopf bzw. Pfeilschwanz, welche ebenfalls individuell definiert werden können. Standardmäßig ist dabei der Pfeilkopf als normaler Pfeil gesetzt. Die Kante eines gerichteten Graphen wird dabei durch „->“ definiert. Bei einem „graph“ wird die Kante durch „--“ definiert und in diesem Fall können keine Modifikationen der Kantenenden vorgenommen werden. In Abbildung 8 sind ein ungerichteter und ein gerichteter Graph mit dem jeweiligen Quellcode dargestellt.

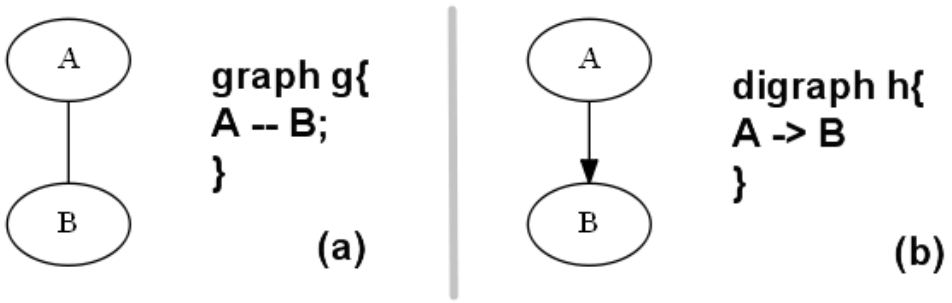


Abbildung 8 Erstellte Graphen mit GraphViz, (a) zeigt das Beispiel eines ungerichteten Graphen und (b) eines gerichteten Graphen.

Für die Erstellung eines Graphen stellt GraphViz verschiedene Renderer zur Verfügung, welche unterschiedliche Graphen generieren und auch mit einer unterschiedlichen Anzahl an Knoten arbeiten können. Als einfachste Möglichkeit der Zeichnung steht der Renderer „dot“ zur Verfügung. Dieser zeichnet den Graphen in verschiedene hierarchische Ebenen, wobei

eine Kante immer Knoten aus verschiedenen Ebenen miteinander verbindet. Darauffolgend ist der Renderer „neato“ verwendbar. Dieser Renderer dient zur Erstellung von ungerichteten Graphen, welcher ebenfalls hierarchische Strukturen besitzt. Zusätzlich können durch die Nutzung von physikalischen Modellen Anziehungs- und Abstoßungkräfte zwischen den einzelnen Knoten simuliert werden und somit der Graph dynamischer gestaltet werden [Kamada & Kawai, 1989].Ein weiterer Renderer ist „fdp“, welcher ebenfalls physikalische Modelle für die Erstellung des Graphen nutzt und ebenfalls für ungerichtete Graphen verwendet werden kann. Neben den hier genannten Renderern existieren noch weitere. Abbildung 9 zeigt die Darstellung eines Graphen mittels der drei genannten Renderer.

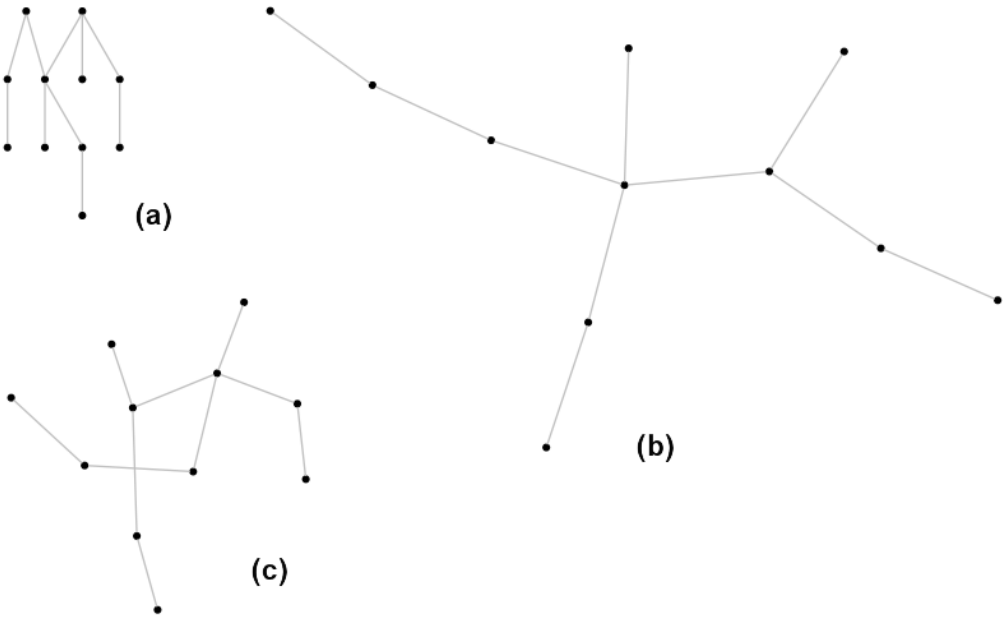


Abbildung 9 Auswahl von GraphViz Renderern. Alle Graphen beruhen auf den gleichen Kanten und Knoten, werden jedoch durch die Renderer unterschiedlich dargestellt. Dabei nutzt (a) dot, (b) neato und (c) fdp.

GraphViz benötigt für die Erstellung eines Graphen einen Zielort für die Speicherung. Dies macht es für webbasierte Anwendungen weniger interessant als beispielsweise die alternative JavaScript Bibliothek D3. Jedoch kann durch die Nutzung des JavaScripts „viz.js“ [URL-5] GraphViz auch für webbasierte Anwendungen genutzt werden.

2 Methoden

Da die CyanoFactory Knowledge Base für alle Projektpartner zur Verfügung stehen soll, ist sie als Webapplikation ausgearbeitet worden. Somit werden auch die Applikationen hinsichtlich der Visualisierung so gestaltet, dass diese für die Nutzung über das Internet geeignet sind. Aufgrund dieser Auflage wurde mittels Python und Django gearbeitet und somit die Methoden und Funktionen für den Aufbau der einzelnen Visualisierungen und Funktionen innerhalb der Knowledge Base erstellt. Es wurden dabei zwei Funktionen in die Knowledge Base integriert. Zum einen **CyanoInteraction** und zum anderen **CyanoDesign**. **CyanoInteraction** stellt dabei die Interaktionen zwischen Proteinen, sowie zwischen Proteinen und Chemikalien dar. **CyanoDesign** wird dazu genutzt, metabolische Abläufe zu berechnen und stellt diese gleichzeitig auch dar.

2.1 CyanoInteraction

Mittels des Tools **CyanoInteraction** sollten die Interaktionen eines ausgewählten Proteins dargestellt werden. Dabei standen die Interaktionen der Proteine im Vordergrund. Als Datengrundlage diente dabei die STRING Datenbank Version 9.1 und die STITCH Datenbank Version 4. In Abbildung 10 sind die verwendeten Teile der Datenbankschemata der STRING Datenbank und der STITCH Datenbank dargestellt. Die Informationen wurden dabei mittels Python aus der PostgreSQL Datenbank ausgelesen und mit JavaScript visualisiert.

Da CyanoFactory Knowledge Base eine Webanwendung ist, gilt dies auch für **CyanoInteraction** als Bestandteil der Knowledge Base. Für die Abfrage einer Interaktion wird daher vom Nutzer eine Anfrage zu einem bestimmten Protein gestellt. Diese Anfrage wird von der Knowledge Base verarbeitet und gibt als Antwort eine dynamisch generierte HTML-Seite mit spezifischem Inhalt zu den Interaktionen des gewünschten Proteins zurück.

Das daraus resultierende Netzwerk an Interaktionen wird als ungerichteter Graph aufgefasst. Die Proteine innerhalb einer Interaktion werden dabei als Knoten und die Interaktionen zwischen zwei Proteinen als Kanten des Graphen gesehen. Ein Knoten innerhalb dieses Netzwerkes beinhaltet dabei wichtige Informationen für die spätere Darstellung des jeweiligen Proteins. Diese zusätzlichen Informationen bzw. Attribute oder auch Eigenschaften umfassen beispielsweise den Namen des Proteins, dessen ID und die Anzahl an benachbarten Proteinen. Auch die Kanten beinhalten weitere Attribute für die spätere Darstellung. Zu diesen Informationen zählt unter anderem der Interaktionsscore der verbundenen Proteine, sowie die einzelnen Wertungen aus denen sich der Interaktionsscore ergibt. Die Informationen, welche Knoten und Kanten beinhalten, ist in Tabelle 3 dargestellt. Neben der Interaktion zwischen Proteinen wird auch die Interaktion von Proteinen mit Chemikalien beschrieben.

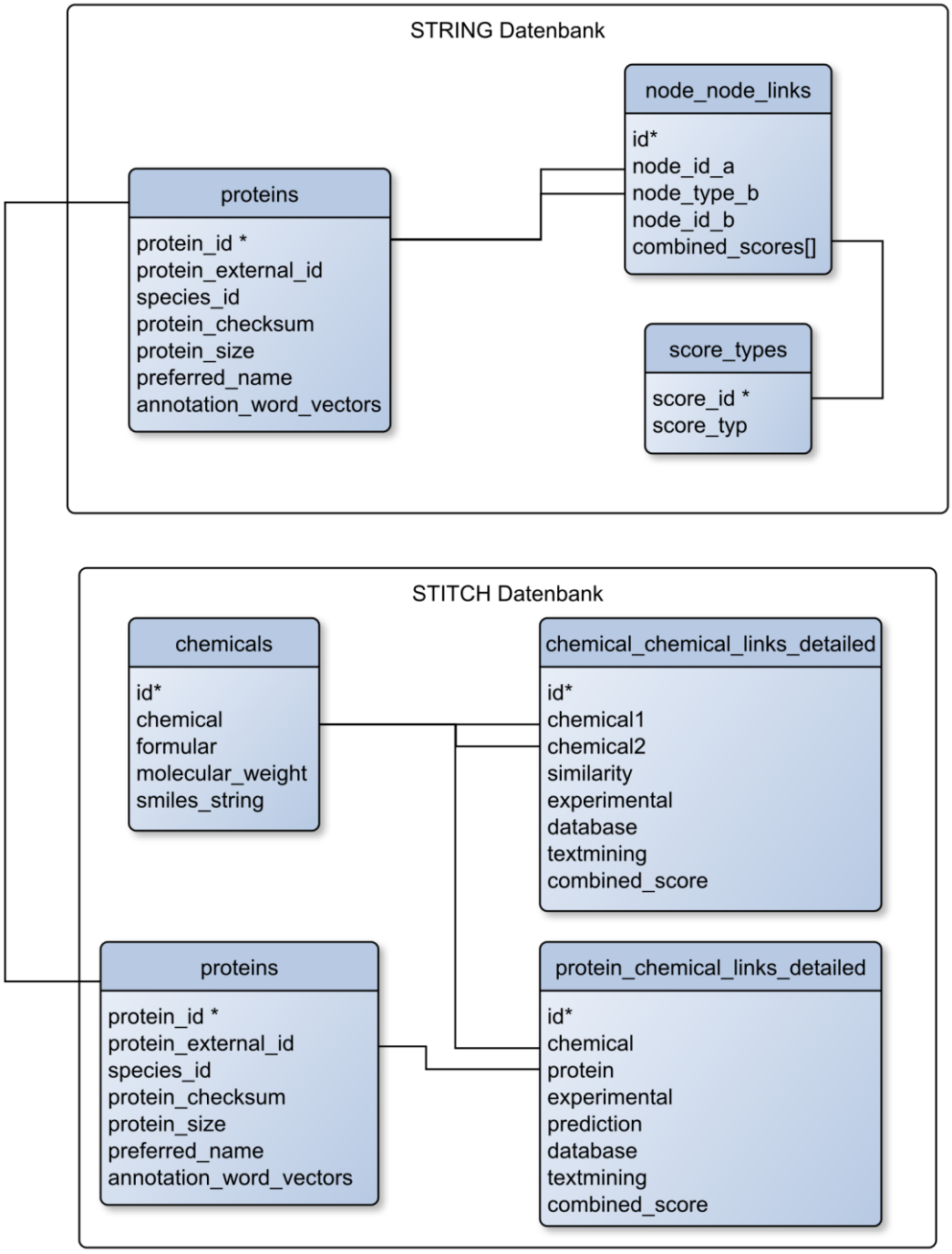


Abbildung 10 Datenbankschema der genutzten Teile von STRING und STITCH. Die jeweiligen Primärschlüssel sind mit einem * versehen. Da die STITCH Datenbank auf der STRING Datenbank aufbaut, sind die Tabellen `Proteins` beider Tabellen gleich aufgebaut. Die STITCH Datenbank nutz jedoch die Informationen von STRING Version 9.05.

Tabelle 3 Knoten- und Kanteneigenschaften. Aufteilung in Eigenschaften welche nur in Kanten bzw. Knoten vorkommen. Für jede Eigenschaft gibt es neben dem Namen die jeweilige Beschreibung.

Knoten

Information	Beschreibung
Protein ID	ID des Protein nach STRING, mit vorangestellten „P“
Name	Proteinname
GeneID	Externe ID nach STRING
Annotation	Annotation des Proteins nach String
Protein	Angabe ob der Knoten ein Protein ist (1) oder nicht (0)
Hood	Anzahl an benachbarten Proteinen

Kanten

Information	Beschreibung
Score	Interaktionswert laut STRING
Homology	Homology Score laut STRING
Experiment	Experimenteller Score laut STRING
Database	Datenbank Score laut STRING
Textmining	Textmining Score laut STRING
Genfusion	Genfusions Score laut STRING
Coocurence	Coocurence Score laut STRING
Neighborhood	Nachbarschafts Score laut STRING
Coexpression	Coexpressions Score laut STRING

2.1.1 Erzeugung eines Protein-Protein-Interaktionsgraphen

Bei der Erzeugung eines Interaktionsgraphen ist n, die Anzahl der zu betrachtenden benachbarten bzw. interagierenden Proteine, wichtig. Zu beachten ist, dass $n \geq 0$. Je mehr Interaktionen betrachtet werden, desto mehr Zeit wird für die Generierung des Graphen in Anspruch genommen. Ausgangselement des Interaktionsgraphen W ist dabei das zu betrachtende Protein, dessen Interaktionen aufgedeckt werden sollen. Dieser Interaktionsgraph W ist dabei definiert als $W = (A, E)$, wobei A die Menge aller Knoten (Proteine) und E die Menge aller Kanten (Interaktionen) ist. Jeder Knoten in P besitzt dabei mehrere Eigenschaften. Diese Eigenschaften werden durch den Eintrag des jeweiligen Proteins in der Tabelle `PROTEINS` der STRING Datenbank definiert. Die einzelnen Eigenschaften der Kanten werden durch die Tabelle `NodeNodeLinks` aus der STRING Datenbank bestimmt.

A ist durch die Einträge des ausgewählten Proteins P_0 in der Tabelle `NodeNodeLinks` und n definiert. Es sei a, die Anzahl an Einträgen von P_0 in der Tabelle `NodeNodeLinks`. So ist $|A|$, die Anzahl an Knoten in A, wie in (2.1) definiert.

$$|A| = \begin{cases} a+1 \ : \ n > a \\ n+1 \ : \ n < a \\ 1 : a = 0 \end{cases} \tag{2.1}$$

Die Einträge der Tabelle `NodeNodeLinks` sind dabei absteigend nach dem `combined_Score` sortiert, um die ersten n besten Interaktionen zu ermitteln.

Die Menge der Kanten E ist durch die Anzahl an Interaktionen zwischen den Knoten aus A definiert. Dabei werden nicht nur Interaktionen zwischen P_0 und P_i betrachtet, wobei $i = 1, \dots, (|A| - 1)$ ist. Es werden auch die Interaktionen zwischen P_i und P_j betrachtet, wobei $j = 1, \dots, (|A| - 1)$ und $i \neq j$ ist. Sollte E eine Kante $\{P_i, P_j\}$ beinhalten, so gibt es keine Kante $\{P_j, P_i\}$ in E, da beide die gleichen Eigenschaften besitzen. Aus demselben Grund gibt es keine Kanten $\{P_i, P_0\}$.

Innerhalb des Graphen W werden die Knoten $P_{\max}$ und $P_{\min}$ gesucht. Diese besitzen die maximale bzw. minimale Anzahl an Interaktionspartnern innerhalb der STRING Datenbank. Des Weiteren wird aus E die Kante $e_{\min}$ mit dem geringsten Interaktionswert bestimmt, sowie $e_{\max}$ die Kante mit dem höchsten Interaktionswert. Dies dient der späteren Visualisierung.

2.1.2 Erzeugung eines Protein-Chemikalien-Interaktionsgraphen

Die Erzeugung eines Interaktionsgraphen X von Proteinen mit Chemikalien ähnelt dem Protein-Protein-Interaktionsgraphen W aus Abschnitt 2.1.2. Hierbei ist der bestimmende Faktor l, die Anzahl der zu betrachtenden interagierenden Chemikalien. Dabei ist $l \geq 0$. B ist als Menge der Knoten anzusehen und F die Menge der Kanten. So das gilt, $X = (B, F)$ und die Mächtigkeit von B ist durch (2.2) definiert. Dabei ist b die Anzahl an Einträgen in der Tabelle `ProteinChemicalLinks`, welche einen Treffer auf das Protein P_0 aufweisen. Diese Einträge von `ProteinChemicalLinks` sind dabei wieder absteigend nach dem `combined_Score` zu ordnen.

$$|B| = \begin{cases} b+1 : \ l > b \\ l+1 \ : \ l < b \\ 1 : b = 0 \end{cases} \tag{2.2}$$

F beinhaltet alle Interaktionen zwischen dem Protein P_0 und den interagierenden Chemikalien C_i, wobei $i = 1, \dots, (|B| - 1)$. Des Weiteren beinhaltet F auch alle Interaktionen zwischen den einzelnen Chemikalien $\{C_i, C_j\}$, dabei ist $j = 1, \dots, (|B| - 1)$ und i$\neq j$. F beinhaltet nicht die Kanten $\{C_j, C_i\}$ da diese Kante äquivalent zu $\{C_i; C_j\}$ wären. Aus dem gleichen Grund entfällt

die Kante $\{C_i, P_0\}$ Diese Interaktionen sind durch die Tabelle `ChemicalChemicalLinks` definiert.

2.1.3 Erzeugung eines Chemikalien-Protein-Interaktionsgraphen

Der Graph Y, welcher die Interaktionen einer gegebenen Chemikalie mit Proteinen darstellt ähnelt dem Graph W aus Abschnitt 2.1.1 und dem Graph X aus Abschnitt 2.1.2. Y ist dabei definiert als $Y = (C, G)$, wobei C die Menge der Knoten und G die Menge der Kanten ist. Wichtig bei der Erstellung dieses Graphen ist k, die Anzahl an zu betrachtenden interagierenden Proteinen. Dabei gilt $k \geq 0$ Die Anzahl an Knoten in C ist durch (2.3) definiert. Dabei ist c die Anzahl an Interaktionen welche die Ausgangschemikalie C_0 laut der Tabelle `ProteinChemicalLinks` mit Proteinen eingeht.

$$|C| = \begin{cases} c + 1 : l > c \\ k + 1 : l < c \\ 1 : c = 0 \end{cases} \quad (2.3)$$

G beinhaltet alle Interaktionen zwischen der Chemikalie C_0 und den interagierenden Proteinen P_i, wobei $i = 1, \dots, (|C| - 1)$. Des Weiteren beinhaltet G auch alle Interaktionen zwischen den einzelnen Proteinen $\{P_i, P_j\}$, dabei ist $j = 1, \dots, (|C| - 1)$ und i≠ j. F beinhaltet nicht die Kanten $\{P_j, P_i\}$ da dies dieselben Kanten wären wie $\{P_i; P_j\}$. Aus dem gleichen Grund entfällt die Kante $\{P_i, C_0\}$ Diese Interaktionen sind durch die Tabelle `NodeNodeLinks` definiert.

2.1.4 Erzeugung eines Chemikalien-Chemikalien-Interaktionsgraphen

Graph Z, stellt die Interaktionen zwischen den einzelnen Chemikalien dar und ähnelt den Graphen aus den vorangegangen Abschnitten 2.1.1, 2.1.2 und 2.1.3. Z ist dabei definiert durch $Z = (D, H)$, wobei D die Menge der Knoten und H die Menge der Kanten ist. Relevant für die Erstellung des Graphen Z ist m, die Anzahl an zu betrachtenden interagierenden Chemikalien. Dabei gilt $m \geq 0$. Die Anzahl an Knoten in C ist durch (2.4) definiert. d ist dabei die Anzahl an Interaktionen welche die Ausgangschemikalie C_0 laut der Tabelle `ChemicalChemicalLinksDetailed` mit anderen Chemikalien eingeht.

$$|D| = \begin{cases} d + 1 : l > d \\ m + 1 : l < d \\ 1 : d = 0 \end{cases} \quad (2.4)$$

H beinhaltet alle Interaktionen zwischen der Chemikalie C_0 und den interagierenden anderen Chemikalien C_i, wobei $i = 1, \dots, (|D| - 1)$. H enthält auch alle Interaktionen zwischen den einzelnen Chemikalien $\{C_i, C_j\}$, dabei ist $j = 1, \dots, (|D| - 1)$ und i≠ j. H beinhaltet nicht die Kante $\{C_j, C_i\}$ da diese Kante äquivalent zu $\{C_i; C_j\}$ wäre. Aus dem gleichen Grund entfällt die Kante $\{C_i, C_0\}$ Diese Interaktionen sind durch die Tabelle `ChemicalChemicalLinksDetailed` definiert.

2.1.5 Vereinigung von Interaktionsgraphen

Die beschriebenen Graphen in den vorangestellten Abschnitten 2.1.1 - 2.1.4 lassen sich teilweise miteinander vereinen bzw. durch die dargelegte Vorgehensweise erweitern. Eine Vereinigung der Graphen W und X, Protein-Protein-Interaktion und Protein-Chemikalien-Interaktion, ist dabei sinnvoll. Eine weitere sinnvolle Vereinigung ist die der Graphen Y und Z, also der Chemikalien-Protein-Interaktion und der Chemikalien-Chemikalien-Interaktion. Dabei entsteht zum einen der Graph $S = (A \cup B, E \cup F)$, ein Graph der die Interaktion eines Proteins mit Proteinen und Chemikalien beinhaltet. Zum anderen entsteht der Graph $R = (C \cup D, G \cup H)$. Dieser beinhaltet die Interaktionen einer Chemikalie mit Proteinen und Chemikalien.

Durch die dargelegte Vorgehensweise für die Erstellung der einzelnen Graphen W, X, Y und Z, lassen sich die jeweiligen Graphen durch die Anwendung der Regeln für die Erstellung der jeweils anderen erweitern. Zum Beispiel kann der Graph W für die Protein-Protein-Interaktion so erweitert werden, dass für alle beteiligten Proteine P_i eine definierte Anzahl an interagierenden Chemikalien hinzugefügt werden kann. Dies ist anhand der Vorschrift für die Erstellung des Graphen X möglich.

2.1.6 Visualisierung des Graphen

Für die webbasierte Darstellung der Interaktionsgraphen wird JavaScript mit der Bibliothek D3 und dem darin enthaltenen Force Layout genutzt. Dazu werden die an den Graphen übergebenen Eigenschaften genutzt. Mittels dieser wird die Größe und Form der Knoten, die Länge der Kanten sowie die Farbe der Knoten und Kanten definiert.

Die in den folgenden Abschnitten verwendete Bezeichnung des Graphen, steht mit keiner Bildungsvorschrift eines Graphen oder deren genutzten Variablen aus den vorherigen Abschnitten 2.1.1 - 2.1.4 in Zusammenhang. Der zu visualisierende Graph wird daher nur als Graph G bezeichnet und besteht aus $G = (V, E)$, wobei V die Menge der Knoten und E die Menge der Kanten ist.

2.1.6.1 Definition von Knotengröße und –form

Durch die dynamisch variierende Menge an Knoten V des betrachteten Graphen G ist es nötig die Darstellungsgröße ebenfalls dynamisch zu gestalten. r, der Radius des Kreises welcher die Größe eines Knoten definiert, wird dabei wie in (2.5) definiert. Dabei ist $|V|$ die Anzahl der Knoten in der Menge der Knoten V des aktuellen Graphen.

$$r = \frac{1}{|V| * a} \tag{2.5}$$

a ist eine Konstante welche sich aus der Höhe h und der Breite w der für die Visualisierung des Graphen zur Verfügung stehenden Fläche ergibt. Diese Konstante wird wie folgt berechnet:

$$b = \begin{cases} \frac{w}{h} : w \geq h \\ \frac{h}{w} : w < h \end{cases} \quad (2.6)$$

$$a = b * 4 \quad (2.7)$$

Die Form des Knotens hängt von seiner Art ab, wie in Abbildung 11 dargestellt ist. Handelt es sich bei dem Knoten um ein Protein, so wird dieser als gefüllter Kreis dargestellt. Ist der Knoten dagegen eine Chemikalie, so wird dieser als Ring visualisiert.

Abbildung 11 Darstellung von Protein und Chemikalie in Interaktionsnetzwerk

2.1.6.2 Definition von Kantenlänge und –breite

Die Länge einer jeden Kante ist individuell zu bestimmen. Dabei sind Knoten mit einem hohen Interaktionswert dichter beieinander als zwei Knoten mit einem niedrigeren Interaktionswert. Die Länge der Kante $e \in E$ wird als kl_e bezeichnet. kl_e berechnet sich aus dem Interaktionswert s_e der Kante e und dem Mittelwert des Interaktionswerts aller Kanten aus E und wird wie in (2.8) bestimmt.

$$kl_e = \frac{\frac{1}{s_e} * \frac{\sum_{i \in E} s_i}{|E|} * k_1}{e^{|V|/k_2}} \quad (2.8)$$

k_1 und k_2 werden dabei wie folgt definiert:

$$k_1 = b * 8 \quad (2.9)$$

$$k_2 = b * 2 \quad (2.10)$$

Die Kantenbreite kb wird nicht für jede Kante einzeln definiert, sondern für alle Kanten von E und ist abhängig von der Anzahl an Knoten in V. Sie berechnet sich wie in (2.11) beschrieben.

$$kb = \frac{1}{|V| * b} \tag{2.11}$$

2.1.6.3 Definition von Knoten- und Kantenfarben

Für die visuelle Darstellung der Knoten- und Kanteninformationen wird der Farbverlauf von Knoten und Kanten im Vorfeld bestimmt. Dafür werden der maximale und minimale Nachbarschaftswert sowie der maximale und minimale Interaktionswert benötigt.

Der Farbverlauf für die Knoten und Kanten ist dabei jeweils mit vier Farben definiert. Diese Farben sind jeweils rot, gelb, grün und blau. Die Knoten besitzen zusätzlich die Farbe schwarz, für die Darstellung des Knotens P_0 bzw. C_0. Die Bedeutung der Farben von Knoten und Kanten unterscheidet sich dabei. Der Farbverlauf einer Kante erfolgt dabei von rot, dem minimalen Interaktionswert, über gelb und grün, zu blau, dem maximalen Interaktionswert. Bei einem Knoten ist jedoch der Verlauf minimale Anzahl an Nachbarn zu maximale Anzahl von Nachbarn von blau über grün und gelb zu rot. Die Zwischenwerte für die Farben grün und gelb ergeben sich aus der Spanne zwischen den jeweiligen Minimalwerten und Maximalwerten. (2.12) zeigt dies am Beispiel der Kante.

$$\begin{aligned} rot &: s_{\min} \\ gelb &: s_{\min} + \frac{s_{\max} - s_{\min}}{4} \\ grün &: s_{\max} - \frac{s_{\max} - s_{\min}}{4} \\ blau &: s_{\max} \end{aligned} \tag{2.12}$$

2.1.6.4 Definition des physikalischen Verhaltens des Graphen

Die Nutzung eines physikalischen Modelles für die Visualisierung eines Graphen ermöglicht die interaktive Nutzung dieses Graphen. D3 bietet dazu die Möglichkeit. Dafür müssen verschiedene Parameter des Modelles definiert werden. Diese Parameter sind mit einer kurzen Beschreibung in Tabelle 4 aufgeführt.

Tabelle 4 Parameter für die Visualisierung mittels D3. Teilweise handelt es sich dabei um feste Parameter, andererseits werden die Parameter dynamisch dem gegebenen Graphen entsprechend angepasst.

Parameter	Beschreibung	Wert
gravity	Anziehungskräfte zwischen einzelnen Knoten	0,9
friction	Koeffizient für die Dauer bis Erstarren der Bewegungssimulation	0,1
nodes	Knoten, welche bei der Visualisierung beachtet werden sollen	Alle Knoten aus V
links	Kanten welche bei der Visualisierung beachtet werden sollen	Alle Kanten aus E
linkDistance	Länge der einzelnen Kanten	wie in Abschnitt 2.1.6.2 beschrieben
size	Fläche welche das Modell einnehmen darf – [Breite, Höhe]	[800,400]
charge	An-/Abstoßungskraft einzelner Knoten	Berechnung laut (2.13)

Die An- bzw. Abstoßungskräfte ck der Knoten berechnen sich wie folgt:

$$ck = -10 * |V| \tag{2.13}$$

Anhand der gegebenen Parameter werden jeden Zeitschritt Berechnungen für die Position der Knoten und Kanten durchgeführt. Dabei wird darauf geachtet, dass zwei oder mehrere Knoten nicht ein und denselben Platz einnehmen können, sondern immer ein definierter Abstand von $2r$ vorliegt. Diese Abstandsberechnung basiert auf dem Quadtree-Algorithmus und ist Bestandteil von D3 [URL-6].

2.1.7 Funktionsweise CyanoInteraction

Die Funktionsweise bzw. der Ablauf von **CyanoInteraction** ist in Abbildung 12 schematisch dargestellt. Zu Beginn wählt der Nutzer aus, welche Art von Interaktion betrachtet werden soll. Entweder wird ein Protein oder eine Chemikalie als Ausgangspunkt für den Graphen gewählt. Sollte ein Protein Ausgangspunkt sein, so werden serverseitig die einzelnen Subgraphen berechnet. Zuerst wird dabei der Protein-Protein-Interaktionsgraph P erstellt, gefolgt von der Erstellung des Protein-Chemikalien-Interaktionsgraphen C_i für jedes Protein i aus P. Anschließen werden die einzelnen Graphen vereinigt und der Graph für den Anwender visualisiert.

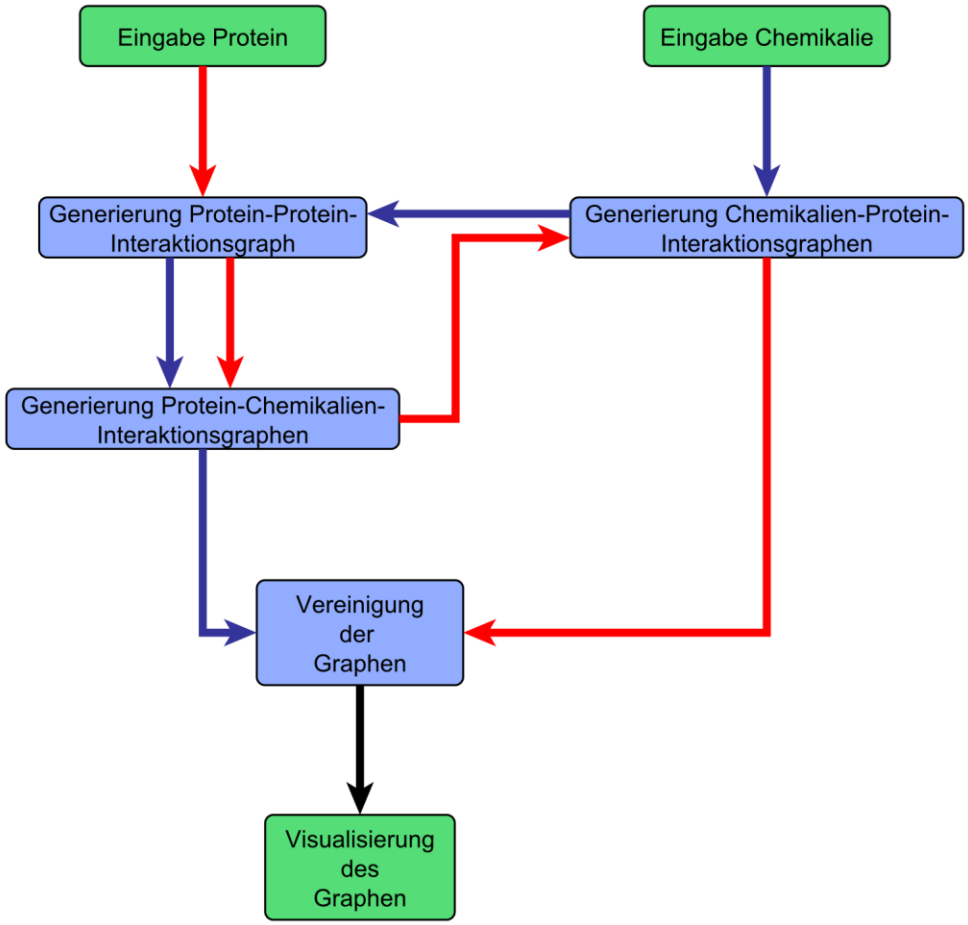


Abbildung 12 Funktionsweise CyanoInteraction, dabei stellen die roten Pfeile den Prozess für die Visualisierung von Protein-Eingaben dar. Blaue Pfeile zeigen den Prozess für die Erstellung eines Interaktionsgraphen beruhend auf der Eingabe einer Chemikalie. Die grünen Boxen stellen Prozesse, welche auf dem System des Nutzers laufen und blaue Boxen stellen die Prozesse des Servers dar

2.2 CyanoDesign

Zur besseren Vorbereitung von Experimenten, bietet es sich an diese zuvor zu simulieren. Mittels **CyanoDesign** soll dies in der CyanoFactory KB umgesetzt werden. Dazu wird die von den spanischen Projektpartnern erstellte Pythonbibliothek PyNetMet genutzt.

Auch hier werden wie schon in Abschnitt 2.1 Graphen erstellt. In dieser Anwendung der CyanoFactory KB soll jedoch der Graph zur Visualisierung der Ergebnisse von Flux-Analysen dienen. Im Gegensatz zu der Anwendung **CyanoInteraction** handelt es sich bei den hier generierten Graphen um gerichtete Graphen. Grund dafür ist die Bildung bzw. Verstoffwechselung einzelner Substrate.

2.2.1 Generierung des Flux-Graphen

Der Flux-Graph F wird definiert durch $F = (M, P)$, wobei M der Menge der Metabolite (Edukte, Produkte) und Enzyme entspricht welche in dem Modell für Flux-Berechnungen enthalten sind. P beschreibt die Menge der Reaktionswege, welche die Substrate mit Proteinen verbinden und Proteine mit Produkten. Jede Kante beinhaltet den Flux der jeweiligen beschriebenen Reaktion. M beinhaltet dabei mindestens drei Knoten. Bestehend aus dem Edukt e, dem Enzym r und dem Produkt p. Damit besteht R mindestens aus den zwei Kanten (e, r) und (r, p). Wenn die Flux-Analyse mehr als eine Reaktion enthält ist es möglich, dass es ein Edukt e_i und ein Produkt p_j gibt, so das gilt $e_i = p_j$ und $i \neq j$. Dabei sind $i, j = 1, \dots, n$ und n die Anzahl an Reaktionen in dem betrachteten Flux-Model.

2.2.2 Visualisierung des Flux-Graphen

Für die Visualisierung des Graphen F wird GraphViz genutzt. Für die Anzeige der einzelnen Flux-Werte f_i einer Reaktion i, wobei $i \in R$ der Menge an Reaktionen des Flux-Modelles entspricht, wird eine Darstellung über die Kantenfarbe und Kantenstärke angewandt. Die Kantenstärke und –farbe gilt dabei für alle bei der Reaktion i beteiligten Kanten.

Die Kantenfarbe ist abhängig davon ob f_i positiv oder negativ ist. Bei einem negativen Flux-Wert ist die Kante rot und bei einem positiven Flux-Wert grün dargestellt.

Die Kantenstärke k_i des Flux-Wertes f_i ist festgelegt auf einen Wert zwischen $k_{\min} = 1$ und $k_{\max} = 20$. Der Flux-Wert wird dafür auf einen Wert in diesem Bereich normiert. Die Normierung erfolgt durch die Bestimmung des maximalen Flux-Wertes $f_{\max}$ und des minimalen Flux-Wertes $f_{\min}$ aus den Beträgen von f_i wie in (2.14) und (2.15) definiert.

$$f_{\max} = \max_{r \in R} |f_r| \tag{2.14}$$

$$f_{\min} = \min_{r \in R} |f_r| \tag{2.15}$$

Anhand von $f_{\max}$ und $f_{\min}$ erfolgt die Überführung von f_i nach k_i wie in (2.16) festgelegt.

$$k_i = \frac{|f_i| - f_{\min}}{(f_{\max} - f_{\min})} * (k_{\max} - k_{\min}) + k_{\min} \qquad (2.16)$$

2.2.3 Funktionsweise CyanoDesign

In Abbildung 13 wird die Funktionsweise von **CyanoDesign** schematisch dargestellt. Vom Anwender wird dabei ein zu bearbeitendes bzw. zu analysierendes Modell gewählt. Anschließend wird anhand des gewählten Modelles ein Flux-Graph generiert. Bei diesem Prozess ist der Anwender nicht involviert, da dies ein serverseitiger Prozess ist. Stattdessen legt der Nutzer in der Zwischenzeit die Parameter für die Simulation fest sowie die Referenzreaktion welche von PyNetMet benötigt wird. Ist der Nutzer mit der Parametereinstellung fertig so wird die FBA-Berechnung wieder serverseitig ausgeführt und anhand des Ergebnisses werden die Kanten des Graphen definiert. Nach Abschluss der Berechnung wählt der Nutzer die darzustellende Reaktionen aus und diese werden anschließend ausgegeben.

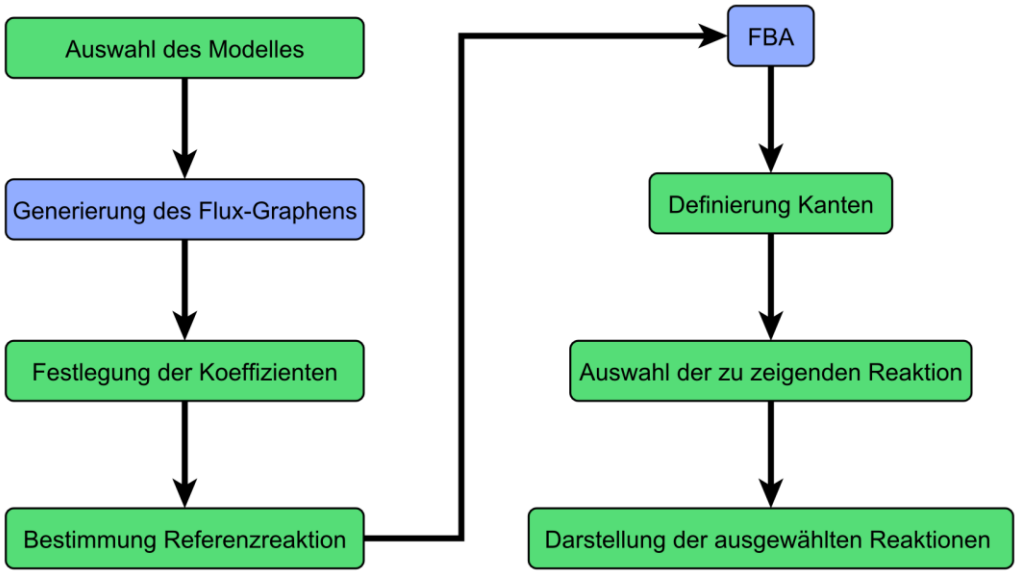


Abbildung 13 Funktionsweise CyanoDesign, dabei sind grün die Prozesse welche auf dem System des Nutzers laufen. Blau dargestellt sind die serverseitigen Prozesse.

3 Ergebnisse und Auswertung

3.1 CyanoInteraction

Die Anwendung **CyanoInteraction** ermöglicht eine gezielte Abfrage zu einem bestimmten Protein bzw. einer bestimmten Chemikalie und deren Interaktionspartnern. Dabei sind als Interaktionspartner sowohl andere Proteine als auch andere Chemikalien zulässig. Die Anzeige der Chemikalien ist optional. Standardmäßig werden bei einer Anfrage, egal ob Protein oder Chemikalie, die zehn interagierenden Proteine mit dem höchsten Interaktionswert dargestellt. Zudem wird zu jedem Protein innerhalb des Interaktionsnetzwerkes die Chemikalie angezeigt, die den höchsten Interaktionswert mit dem jeweiligen Protein bildet. Es wird zwischen allen interagierenden Proteinen und Chemikalien eine Kante gezeichnet. Abbildung 14 zeigt das Ergebnis einer Suche mit dem Protein rpoA, der RNA Polymerase Untereinheit Alpha.

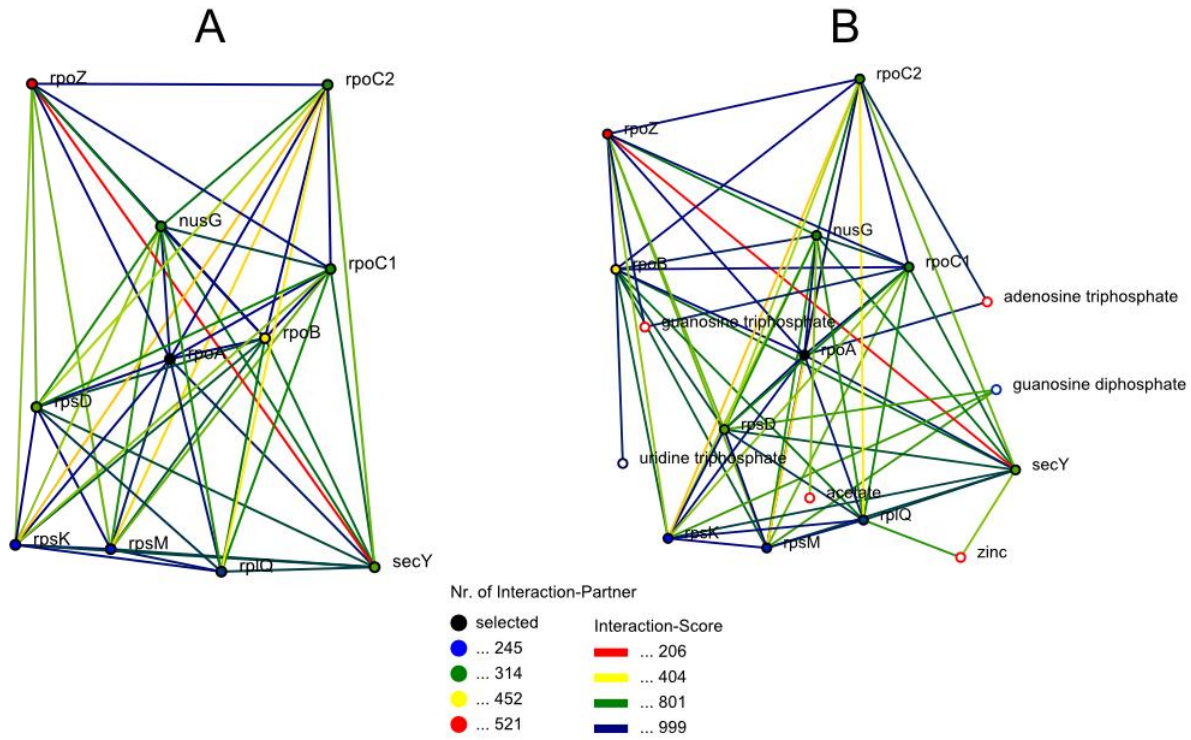


Abbildung 14 Interaktionsnetzwerk von rpoA. A zeigt die Interaktionen der zehn Proteine mit den besten Interaktionsbewertungen zu rpoA und deren Interaktion miteinander. B stellt zusätzlich die Chemikalie mit dem besten Interaktionsscore zu jedem Protein dar.

Die Positionierung der einzelnen Knoten innerhalb des Interaktionsnetzwerkes ist dabei zufällig und variiert bei jedem Aufruf eines Proteins bzw. einer Chemikalie.

Durch die Selektion eines Proteins oder einer Chemikalie, wie in Abbildung 15 zu sehen ist, werden nur die Interaktionspartner dieses Knotens gezeigt bzw. hervorgehoben. Alle nicht mit dem ausgewählten Protein interagierenden Elemente des Graphen werden ausgeblendet. Die

ausgewählten Knoten sind dabei in einer Liste mit zusätzlichen Informationen am Rande des Netzwerkes dargestellt. Durch die Auswahl eines Elementes aus dieser Liste wird dieses zum Ausgangspunkt eines neu generierten Netzwerkes.

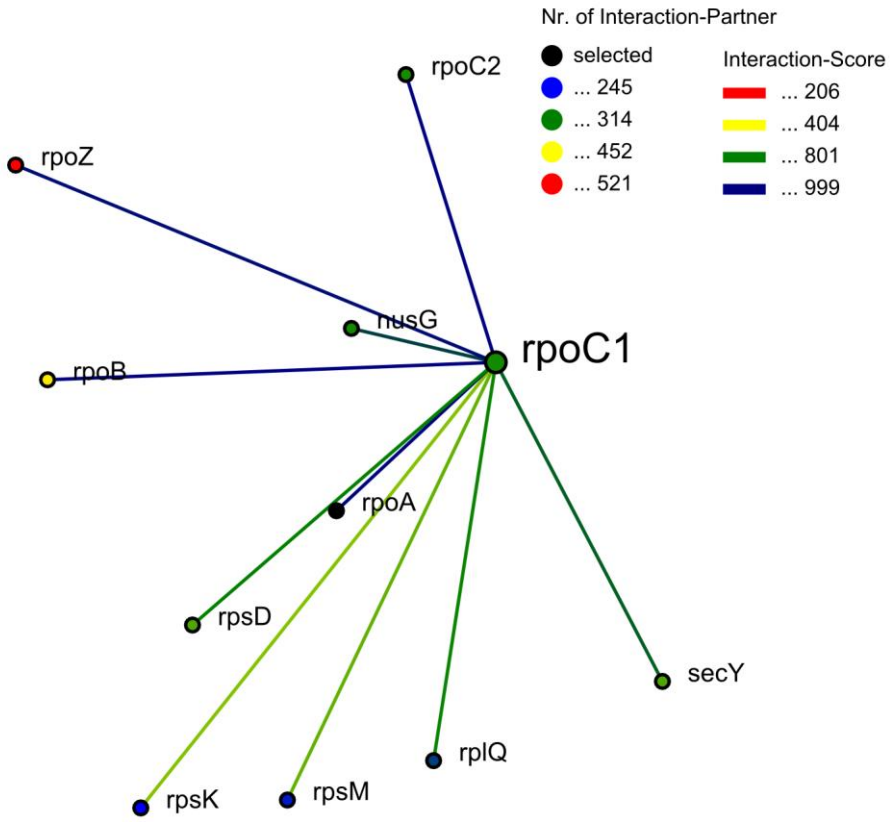


Abbildung 15 Auswahl eines Proteins aus rpoA Interaktionsnetzwerk. Durch die Auswahl werden alle nicht mit dem ausgewählten Proteine rpoC1 interagierende Proteine ausgeblendet.

Das mittels **CyanoInteraction** dargestellte Interaktionsnetzwerk kann dabei nach Belieben erweitert bzw. verringert werden. Die Erweiterungsmöglichkeiten sind dabei variabel. Es können die Anzahl an interagierenden Proteinen oder die Anzahl an interagierenden Chemikalien getrennt voneinander erhöht bzw. verringert werden. Durch die Veränderung der Anzahl an darzustellenden Knoten, ändert sich die Größe aller Knoten, sowie die Länge der einzelnen Kanten. Die Minimierung bzw. Maximierung des Netzwerkes ist in Abbildung 16 dargestellt. Nach der Änderung der Knotenmenge bleiben dabei alle Knoten erkennbar und auswählbar.

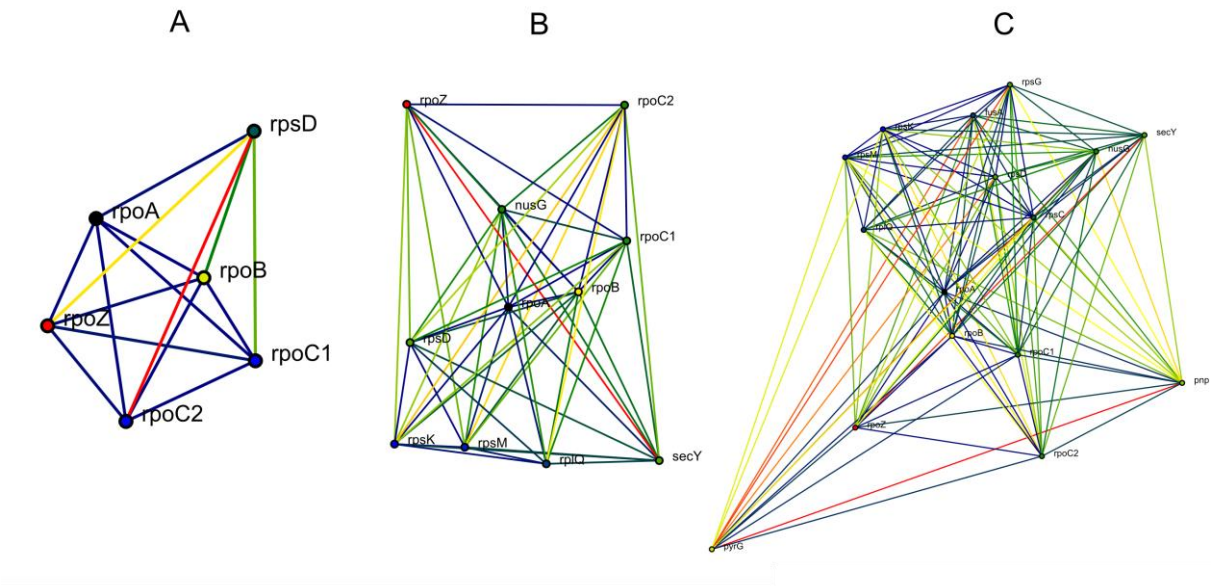


Abbildung 16 Variierung der Anzahl an Interaktionspartnern in Netzwerk. A zeigt ein Netzwerk mit insgesamt sechs Proteinen, B mit 11 Proteinen und C mit 16 Proteinen. Abhängig von der Anzahl an Proteinen und Chemikalien innerhalb des Netzwerkes werden die Knoten und Kanteneigenschaften des Graphen angepasst. Ein kleinerer Graph besitzt größere Knoten und kürzere Kanten als ein großer Graph.

Anhand der gegebenen Kanteneigenschaften ist es möglich, dass eine Auswahl nach dem gewünschten Interaktionstyp erfolgen kann. Die Auswahl wird dabei durch den Nutzer getroffen. Anhand dieser ändern sich zum einen die Kantenfarben, welche über die ausgewählten Interaktionstypen definiert sind. Zum anderen werden die Kanten ausgeblendet die sich nicht über die gewählten Interaktionstypen definieren. Ein Beispiel für die Auswahl einzelner und mehrerer Interaktionstypen ist mit rpoA als Ausgangspunkt für das Interaktionsnetzwerk in Abbildung 17 dargestellt. Die Farbänderung der Kanten ergibt sich durch die Neuausrichtung des maximalen Interaktionswertes $s_{\max}$ und des minimalen Interaktionswertes $s_{\min}$, da sich die jeweiligen Interactionwerte s_i der Kanten $e_i \, \epsilon \, E$ und $i = 1, \dots, |E|$ durch die Auswahl der Interaktionstypen ändern. Der exakte Interaktionswert einer Interaktion zwischen zwei Proteinen ist anhand der Visualisierung und der damit verbundenen Kantenfarbe nicht erkennbar. Die Kantenfarbe dient nur als Richtwert zum Vergleich, wie die einzelnen Interaktionswerte zueinander stehen. Die vier Richtwerte für die Interaktionswerte sowie deren Farben rot, gelb, grün und blau werden bei jeder dargestellten Interaktion, sowie bei der Auswahl von verschiedenen Interaktionstypen angepasst. Die Länge der Kanten bleibt dabei unverändert, so dass diese weiterhin Informationen über die jeweiligen Interaktionswerte zu allen Interaktionstypen geben.

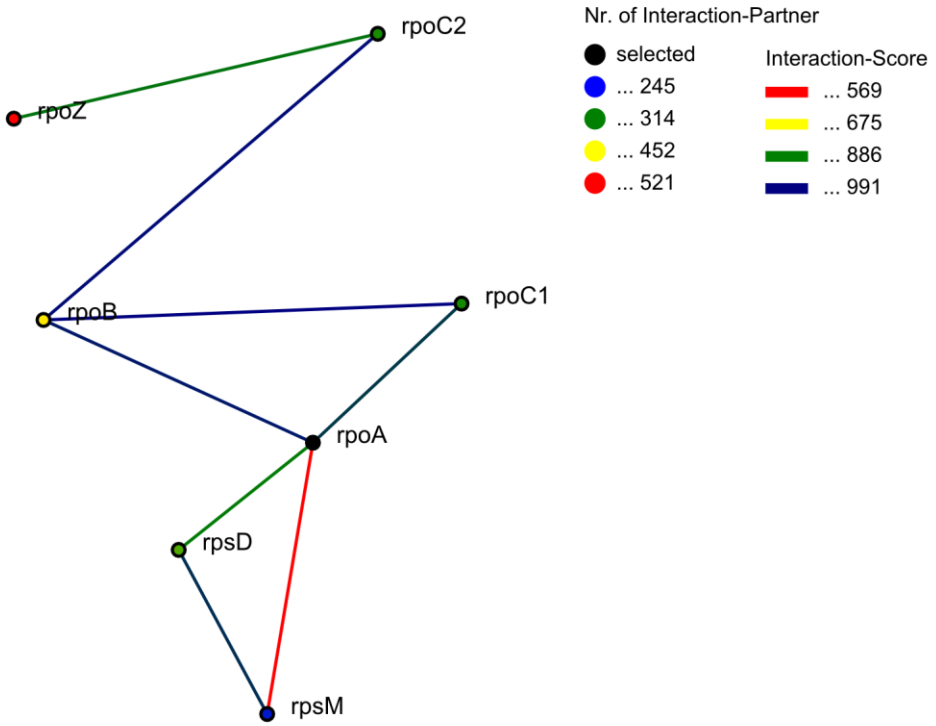


Abbildung 17 Auswahl von Interaktionstypen innerhalb des rpoA Netzwerkes. Es werden nur die Proteine angezeigt, welche durch die definierten Interaktionstypen bzw. Interaktionsquellen miteinander verbunden sind. In dem dargestellten Fall sind dies, Experimente, Textmining und Kookkurrenz. Die Farben der Kanten werden dabei neu berechnet.

Die Positionen der einzelnen Proteine des visualisierten Interaktionsnetzwerkes können nach Belieben durch Drag and Drop des jeweiligen Proteins, wie in Abbildung 18 zu sehen ist, verschoben werden. Durch die Festlegung der Positionen mehrere Proteine besitzen die Kantenlängen keine Informationen mehr über die ursprünglichen Interaktionscores. Auch ohne eine Veränderung der Position können die Längen der Kanten täuschen. Grund dafür ist der hohe Grad an Interaktionen zwischen den einzelnen Bestandteilen des Netzwerkes, dadurch sind die errechneten Kantenlängen aus Abschnitt 2.1.6.2, jeweils nur die Mindestlänge der jeweiligen Kante.

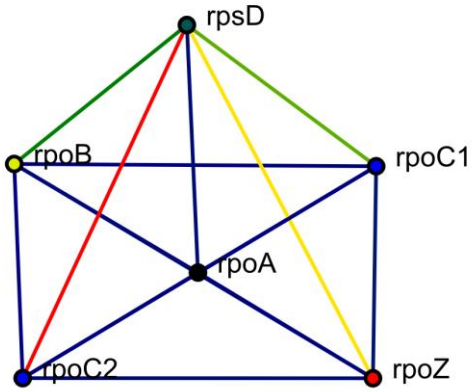


Abbildung 18 Drag and Drop Anordnung von Proteinen durch den Nutzer. Durch die feste Positionierung der Knoten des Graphen, werden die physikalischen Eigenschaften des dahinterliegenden Modelles eingeschränkt. Durch die Definition der Knotenpositionen verliert die Kantenlänge die Information über den Interaktionswert.

Neben der graphischen Darstellung der Interaktionen eines ausgewählten Proteins oder einer Chemikalie, werden zu dessen Interaktionspartner auch der genaue Interaktionswert und die jeweiligen Interaktionstypen wie in Tabelle 5 aufgelistet. Durch die Auswahl eines Eintrages wird dessen Netzwerk aufgerufen. Die Liste ist dabei ebenfalls dynamisch gestaltet und es ist möglich zwischen den Interaktionspartnern Protein und Chemikalien zu wechseln, bzw. können beide gleichzeitig dargestellt werden.

Tabelle 5 Interaktionstabelle von rpoA, dabei kann durch den Nutzer die Auswahl getroffen werden ob sowohl Proteine und Chemikalien angezeigt werden sollen oder nur eines dieser beiden. Mit der Generierung der Seite werden nur die ersten zehn Interakteure mit den höchsten Interaktionswerten angezeigt. Weitere können nachträglich hinzugeladen werden.

ID	Name	Annotation	Homology	Experiment	Database	Textmining	Genfusion	Coocurence	Neighborhood	Coexpression	Score
P16959	rpoB	DNA-directed RNA polymerase subunit beta		•	•	•		•	•	•	999
P18135	rpoC1	DNA-directed RNA polymerase subunit gamma		•	•	•		•	•	•	998
P16960	rpoC2	DNA-directed RNA polymerase subunit beta'		•	•	•			•	•	996
P19064	rpoZ	DNA-directed RNA polymerase subunit omega		•	•	•			•	•	995
P17556	rpsD	30S ribosomal protein S4		•		•	•	•	•	•	990
P16983	rpsK	30S ribosomal protein S11				•	•		•	•	984
P16923	nusG	transcription antitermination protein NusG		•	•	•			•	•	973
P16985	rplQ	50S ribosomal protein L17				•		•	•	•	969
P16980	secY	preprotein translocase subunit SecY				•			•	•	956
P16982	rpsM	30S ribosomal protein S13		•		•		•	•	•	956
P16970	rpsC	30S ribosomal protein S3		•		•			•	•	948
P16411	pnp	polynucleotide phosphorylase/polyadenylase		•	•	•			•	•	942
P16698	pyrG	CTP synthetase			•	•				•	941
P16450	rpsG	30S ribosomal protein S7		•		•			•	•	938

Für die Auswertung von **CyanoInteraction** wurde in Abschnitt 3.1.1 geprüft wie lange die Abfrage der Interaktion von einem Protein mit anderen Proteinen, sowie Chemikalien dauert. In Abschnitt 3.1.1 wurde ebenfalls die Abfragedauer von Interaktionen einer Chemikalie mit Proteinen, sowie Chemikalien durchgeführt. Dabei ist zu beachten, dass die Datenbank mit den Interaktionsinformationen und der Testserver nicht auf dem gleichen System liefen und damit der Datenaustausch über eine drahtlose Netzwerkverbindung vollzogen wurde. Dieses drahtlose Netzwerk unterlag dabei teilweise Verbindungsschwankungen, was die Antwortzeiten der Systeme teilweise verlangsamte. Zusätzlich fanden die Abfragen zu unterschiedlichen Zeitpunkten statt.

3.1.1 Protein-Interaktionen

Es wurde mit allen 3569 Proteinen von *Synechocystis sp. PCC6803*, welche in der STRING Datenbank aufgeführt sind, ein Protein-Protein-Interaktionsgraph P_1 generiert sowie ein Interaktionsgraph P_2 welcher die Interaktion aller Proteine mit deren Chemikalien beschreibt. P_1 besteht dabei aus maximal 11 Proteinen und P_2 aus maximal 11 Proteinen und maximal 11 Chemikalien. Diese wurden wie in 3.1 beschrieben als HTML-Seite mit zusätzlichen Informationen aufgerufen. In diesem Abschnitt wird daher immer die generierte HTML-Seite des jeweiligen Interaktionsgraphen mit P_1 bzw. P_2 bezeichnet. Sollte ein Bezug auf den jeweiligen Interaktionsgraphen genommen werden, so wird explizit darauf hingewiesen.

Es zeigte sich dabei, dass die Abfrage von P_1 mit 2,6 Sekunden durchschnittlich schneller verfügbar war, als die Abfrage von P_2. P_2 benötigte im Durschnitt 4,83 Sekunden. Die Standardabweichung bei dem Aufruf von P_1 betrug dabei 3,29 Sekunden und die Standardabweichung des Aufrufs von P_2 betrug 3,54 Sekunden. Dieser Unterschied zwischen der Abfragedauer von P_1 und P_2 kommt durch die zusätzlich Suche nach Chemikalien, deren Eigenschaften und die Suche nach Interaktionen zwischen den Chemikalien in P_2 zu Stande. Die Analyseergebnisse zu P_1 und P_2 sind in Tabelle 6 dargestellt. Da neben den Graphen auch noch sämtliche Informationen zu den Interaktionen des jeweiligen betrachteten Ausgangsprotein abgerufen werden, kann die Anzahl an Interaktionspartnern in Tabelle 6 höher als die maximale Anzahl an Interaktionspartnern in den jeweiligen Graphen sein. Dabei zeigt sich, das die Anzahl an möglichen Interaktionspartnern von P_2 durchschnittlich fast doppelt so hoch ist wie die durchschnittliche Anzahl an möglichen Interaktionspartnern von P_1. Des Weiteren unterscheiden sich Standardabweichung von den Interaktionspartner-Mittelwert von P_1 und P_2 stark voneinander. Die erhöhte Anzahl an zu betrachtenden bzw. analysierenden Interaktionspartnern, ist für die erhöhte Abrufzeit von P_2 verantwortlich.

Tabelle 6 Ergebnisse der Abfrage von Protein-Protein-Interaktionen mit CyanoInteraction. Es wurden dabei die Abfragezeiten der HTML-Seiten von Protein-Protein Interaktionsgraphen P_1und Abfragezeiten der HTML-Seiten von Protein-Protein/Chemikalien Interaktionsgraphen P_2 betrachtet.

	P_1	P_2
Anzahl betrachteter Proteine	3569	3569
Durchschnittliche Anzahl aufgelisteter Interaktionspartner	127,2	220,15
Standardabweichung aufgelisteter Interaktionspartner	95,67	156,31
Durchschnittliche Aufrufzeit in Sekunden	2,6	4,83
Standardabweichung Aufrufzeit in Sekunden	3,29	3,54
Korrelationskoeffizient zwischen Anzahl Interaktionspartner und Aufrufzeit	0,483	0,899

Dieser Zusammenhang zwischen der Anzahl an Interaktionspartnern und der Abrufzeit der jeweiligen HTML-Seite ist in Abbildung 19 dargestellt.

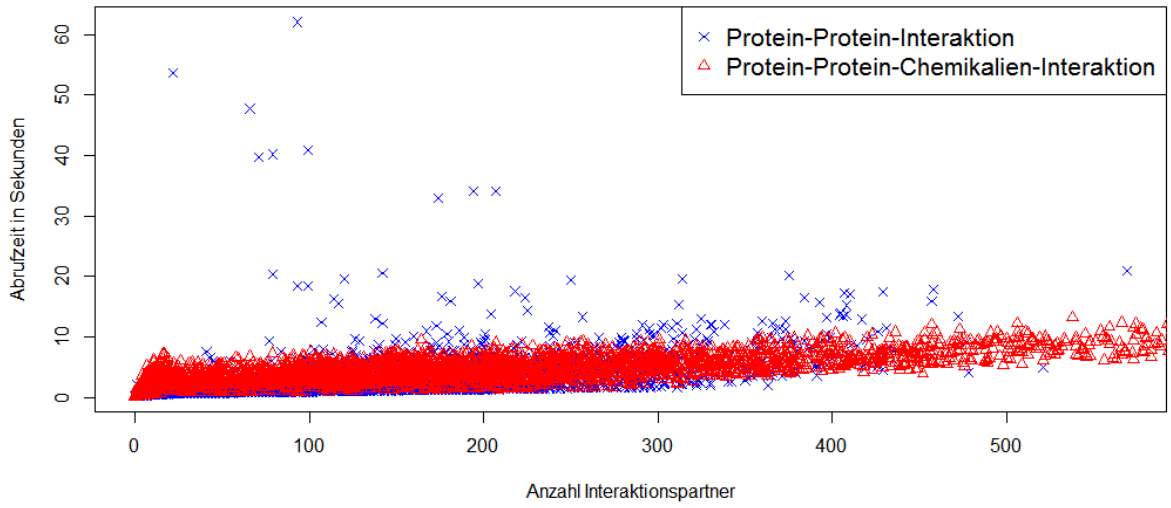


Abbildung 19 Abhängigkeit der Aufrufzeit von der Anzahl an Interaktionspartnern bei Protein Interaktionen. Mit steigender Anzahl an Interaktionspartnern ist eine steigende Abrufzeit zu erkennen.

Während der Korrelationskoeffizient von P_1 zwischen der Anzahl an Interaktionspartnern und der Abrufzeit mit 0,483 einen geringen Zusammenhang zeigt, deutet dieser Korrelationskoeffizient in P_2 mit 0,899 auf einen Zusammenhang dieser hin. Damit zeigt sich, dass nur ein geringer Zusammenhang zwischen der Abfrage von HTML-Seiten mit ausschließlich Protein-Protein-Interaktionen und der Anzahl an Interaktionspartnern besteht. Ein Grund für diesen geringeren Zusammenhang ist der verfügbare Datenraum bzw. die Datenbankeinträge, welche durchsucht werden müssen. Die Anzahl der Interaktionseinträgen

für *Synechocystis sp. PCC6803* in der Tabelle `NodeNodeLinks` der STRING Datenbank beträgt 453966 Diese Anzahl ist im Vergleich zu den Interaktionseinträgen der Tabelle `ProteinChemicalDetailed` mit mehr als 600000 Einträgen und der Tabelle `ChemicalChemicalsDetailed` mit über 11 Millionen Einträgen aus der STITCH Datenbank relativ gering. Und genau diese Einträge der STITCH Datenbank machen die Ausgabe der HTML-Seite von P_2 langsamer. Somit kann durch eine zusätzliche Filterung von Interaktionen mittels eines definierten Interaktionswertes die Anzahl an Interaktionen verringert werden und dadurch der Aufruf einer HTML-Seite beschleunigt werden. Die lokalen Maxima in der Abrufzeit sind auf die Verbindung zwischen Testserver und Datenbank zurückzuführen.

3.1.2 Chemikalien-Interaktionen

Neben den Proteinen als Ausgangspunkt für die Erstellung eines Interaktionsgraphen, wurde dies auch mit Chemikalien durchgeführt. Für die Analyse wurden nur die Chemikalien in der STITCH Datenbank genutzt, welche mit *Synechocystis sp. PCC6803* in Zusammenhang stehen. Aufgrund der hohen Anzahl an interagierenden Proteinen wurde dabei nur eine Stichprobe von 3502 zufälligen Chemikalien getestet. Dazu wurden wieder zwei Interaktionengraphen betrachtet. Zum einen der Interaktionsgraph C_1 mit den Interaktionen der Chemikalie mit Proteinen und der Interaktionsgraph C_2, welcher zu den jeweiligen interagierenden Proteinen nochmals die Interaktionen zu der besten interagierenden Chemikalie bestimmt. Der erstellte Graph besitzt dabei zwischen 1-12 Chemikalien und maximal 11 Proteine. Diese wurden wie in 3.1.1 beschrieben als HTML-Seite mit zusätzlichen Informationen aufgerufen. In diesem Abschnitt wird daher immer die generierte HTML-Seite des jeweiligen Interaktionsgraphen mit C_1 bzw. C bezeichnet. Sollte ein Bezug auf den jeweiligen Interaktionsgraph genommen werden, so wird explizit darauf hingewiesen. Diese wurden wiederum in 3.1 beschrieben als HTML-Seite mit zusätzlichen Informationen aufgerufen. Auch in diesem Abschnitt wird daher immer die generierte HTML-Seite des jeweiligen Interaktionsgraphen mit C_1 bzw. C_2 bezeichnet.

Dabei zeigt sich, wie auch schon in Abschnitt 3.1.1 das der Abruf von C_2 mit 3,77 Sekunden durchschnittlich mehr Zeit benötigte als C_1 mit 1,79. Die Standardabweichung von C_1 betrug dabei 2,1 Sekunden und in C_2 3,98 Sekunden. Sämtliche Ergebnisse von C_1 und C_2 sind in Tabelle 7 zusammengefasst. Es ist auch ein deutlicher Unterschied zwischen der Anzahl an Interaktionspartner zwischen C_1 mit durchschnittlich 20,76 Interaktionspartnern und C_2 mit durchschnittlich 244,62 Interaktionspartnern zu erkennen. Dieser große Unterschied beruht wiederum auf der Anzahl an chemischen Interaktionspartnern für ein Protein innerhalb der STITCH Datenbank, wie schon in Abschnitt 3.1.1 beschrieben.

Tabelle 7 Ergebnisse der Abfrage von Chemikalien-Protein-Interaktionen mit CyanoInteraction. Es wurde dabei die Abfragezeit der HTML-Seiten von Protein-Protein-Interaktionsgraphen P_1und Abfragezeiten der HTML-Seiten von Chemikalien-Protein/Chemikalien Interaktionsgraphen P_2 betrachtet.

	C_1	C_2
Anzahl betrachteter Chemikalien	3502	3502
Durchschnittliche Anzahl aufgelisteter Interaktionspartner	20,76	244,62
Standardabweichung aufgelistete Interaktionspartner	65,79	452,29
Durchschnittliche Aufrufzeit in Sekunden	1,79	3,77
Standardabweichung Aufrufzeit in Sekunden	2,11	3,98
Korrelationskoeffizient zwischen Anzahl Interaktionspartner und Aufrufzeit	0,787	0,607

Betrachtet man jedoch die Abbildung 20 so sieht man, dass trotz der hohen Anzahl an Interaktionspartnern eine relativ geringe Abfragezeit für C_1 und C_2 benötigt wird. Dabei korreliert die Anzahl an Interaktionspartnern mit der Aufrufzeit bei C_1 mit 0,787 besser als die Werte in C_2 mit 0,607. Dies bedeutet, dass trotz der größeren Anzahl an Datenbankeinträgen für Interaktionen von Chemikalien mit Proteinen diese Abfrage weniger Zeit benötigt. Weiterhin ist bei C_1 und C_2 zu erkennen, dass eine hohe Anzahl an erstellten Datenpunkten dicht beieinander liegen.

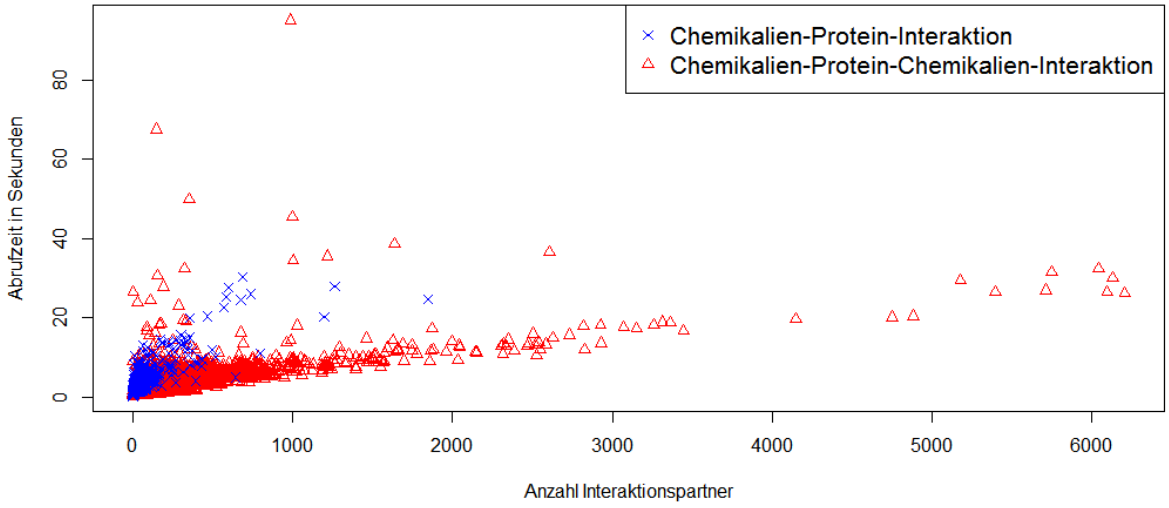


Abbildung 20 Abhängigkeit der Aufrufzeit von der Anzahl an Interaktionspartnern bei Chemikalien Interaktionen. Es zeigt sich das es eine hohe Anzahl an Chemikalien gibt, die nur eine geringe Anzahl an Interaktionen besitzen, wodurch diese schneller dargestellt werden können.

Vergleicht man jedoch die Häufigkeiten der Anzahl an Interaktionspartnern von P_1 und C_1 in Abbildung 21 miteinander, ist ein Unterschied erkennbar. So existieren mehr Chemikalien mit weniger Protein/Chemikalien Interaktionen, als Proteine mit Chemikalien als Interaktionspartner.

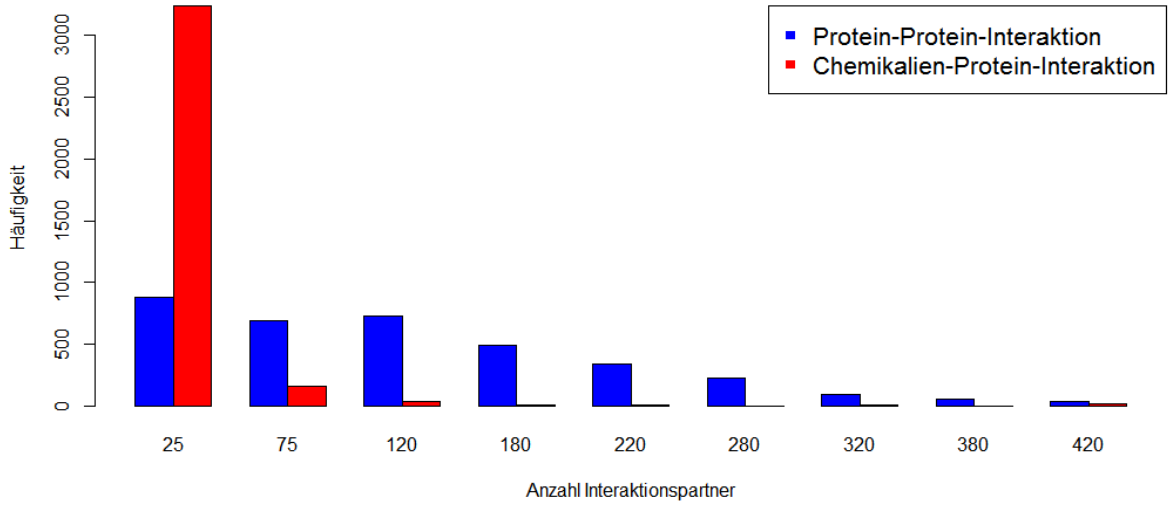


Abbildung 21 Häufigkeiten der Anzahl an Interaktionspartnern in P_1 und C_1, dabei ist ein starker Unterschied zwischen P_1 und C_1 zu beobachten, was zu einer geringeren Aufrufzeit von C_1 führt.

Die Häufigkeit der Anzahl an Interaktionspartnern zwischen den beiden Interaktionstypen unterscheidet sich stark voneinander. Grund für diesen Unterschied ist die zufällige Auswahl an Chemikalien. Die Auswahl an Chemikalien gewährleistet nicht, dass die gewählten Chemikalien eine durchmischte Anzahl an Interaktionspartnern besitzen. Dies zeigt sich in C_1 und C_2 deutlich, da diese viele Interaktionen mit wenigen Interaktionspartnern darstellen. Es zeigt sich somit, dass die Anzahl an gewählten Chemikalien nicht als repräsentativ angesehen werden kann.

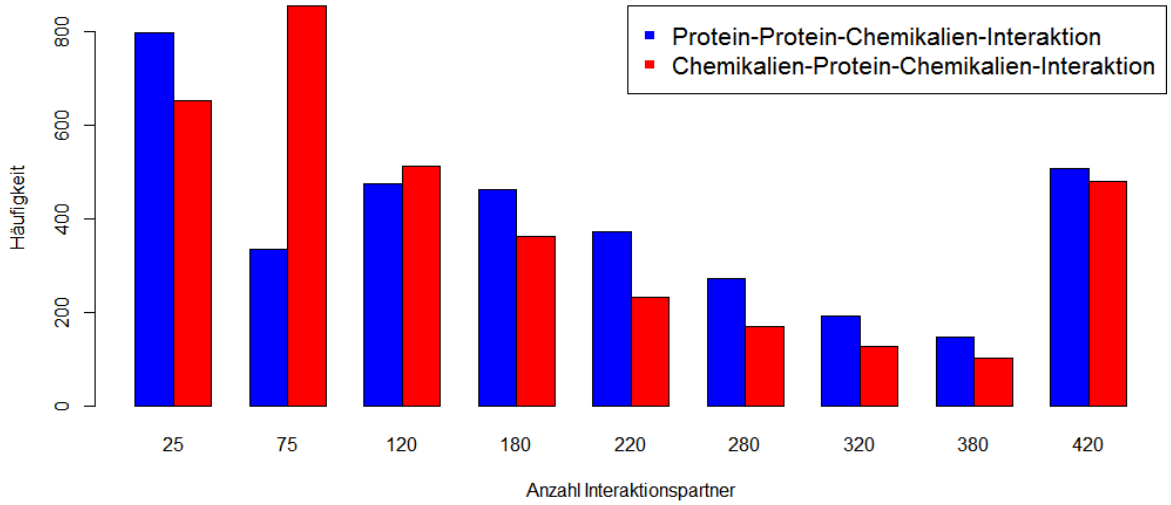


Abbildung 22 Häufigkeiten der Anzahl an Interaktionspartnern in P_2 und C_2. Ähnlich wie bei P_1 und C_1 sind mehr Chemikalien-Protein-Chemikalien-Interaktionen mit einer geringeren Anzahl an Interaktionspartnern vorhanden als Protein-Protein-Chemikalien-Interaktionen

Es wurde deshalb die Anzahl an Chemikalien erhöht. Zusätzlich zu den bereits in C_1 und C_2 erstellten HTML-Seiten wurden zusätzlich rund 5000 weitere erzeugt. Die Ergebnisse der zusätzlichen Analyse werden als C_3 und C_4 bezeichnet und sind in Tabelle 8 dargestellt. C_3 ist dabei die Erweiterung von C_1 und C_4 die Erweiterung von C_2. Dabei zeigt sich, dass durch die zusätzlich getesteten Chemikalien die Aufrufzeit so wie die Anzahl an Interaktionspartnern in C_3 gestiegen ist. Die gewählte Anzahl an Chemikalien ist dabei als Stichprobe nicht optimal, da dies letztendlich nur einen Bruchteil der 453966 in der STITCH Datenbank aufgeführten Chemikalien ist, welche mit *Synechocystis sp. PCC6803* interagieren. Aus zeitlichen Gründen ist jedoch die Betrachtung einer höheren Anzahl an Chemikalien nicht möglich.

Tabelle 8 Vergleich der Chemikalien-Interaktionsergebnisse. Die Erhöhung der betrachteten Chemikalien zeigt dabei starke Auswirkungen auf sämtliche betrachtete Bereiche.

	C_1	C_2	C_3	C_4
Anzahl betrachteter Chemikalien	3502	3502	8500	8500
Durchschnittliche Anzahl aufgelisteter Interaktionspartner	20,76	244,62	77,06	169,25
Standardabweichung aufgelisteter Interaktionspartner	65,79	452,29	167,75	335,24
Durchschnittliche Aufrufzeit in Sekunden	1,79	3,77	2,17	2,97
Standardabweichung Aufrufzeit in Sekunden	2,11	3,98	2,09	3,08
Korrelationskoeffizient zwischen Anzahl Interaktionspartner und Aufrufzeit	0,787	0,607	0,620	0,633

Letztendlich ist wie auch in 3.1.3 die hohe Anzahl an möglichen Interaktionspartnern für die längere Aufrufzeit verantwortlich. Sowohl die Interaktion von Chemikalien mit Proteinen als auch die Interaktion von Chemikalien mit Chemikalien und der damit verbundenen Auflistung, sind Grund für die hohe Abrufzeit. Durch die Einführung eines Schwellwertes für den Interaktionswert, kann die Aufrufzeit gesenkt werden. Abbildung 23 zeigt dies anhand der Ergebnisse von P_1, C_1 und C_3 nochmals.

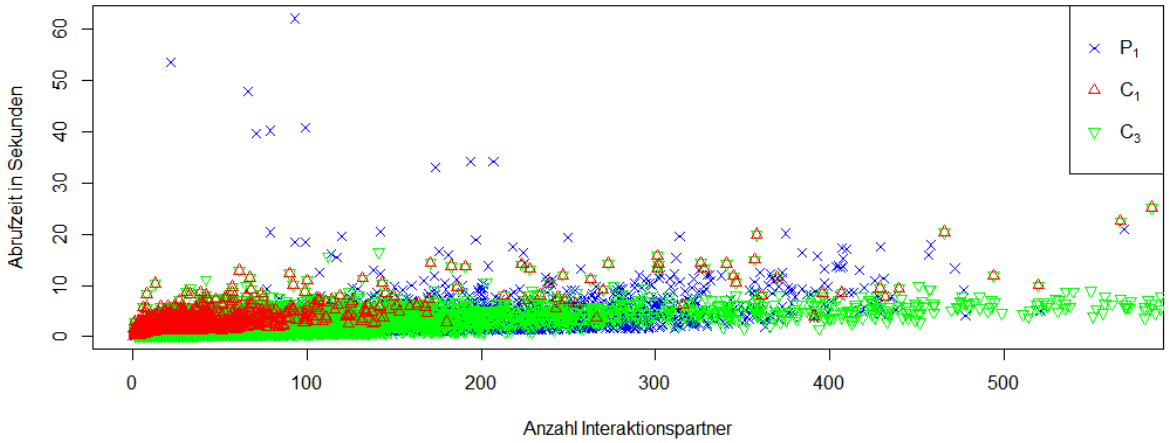


Abbildung 23 Ergebnisse von P_1, C_1 und C_3 im Vergleich. Durch die erhöhte Anzahl an aufgerufenen Chemikalien besitzen P_1 und C_3 ähnliche Verläufe.

3.1.3 Analyse des Protein-Protein-Interaktionsnetzwerkes

Mittels der Bildungsvorschrift von Protein-Protein-Interaktionsgraphen aus 2.1.1 ist es möglich aus den gesamten Einträgen der STRING Datenbank das gesamte bekannte Interaktionsnetzwerk von *Synechocystis sp. PCC6803* darzustellen bzw. zu analysieren.

Das Interaktionsnetzwerk I umfasst 3569 Proteine. Diese sind über 2226983 Interaktionen miteinander verbunden. Die Angabe Interaktion beschreibt dabei nur, dass es einen Interaktionswert zwischen jeweils zwei Proteinen gibt. Dabei besitzen die Proteine im Durchschnitt 127 Interaktionspartner. Die Ergebnisse der Netzwerkanalyse sind in Tabelle 9 aufgelistet. Dabei ist zu erkennen, dass der Interaktionswert im Durschnitt mit 297 über alle Interaktionen von I sehr gering ist. Dies bedeutet, dass innerhalb des Netzwerkes viele unsichere Interaktionen bestehen.

Tabelle 9 Ergebnisse des Interaktionsnetzwerkes I

	Wert
Anzahl Protein	3569
Anzahl Interaktionen	2226983
Durchschnittliche Anzahl an Interaktionspartnern	127
Standardabweichung Anzahl Interaktionspartner	76,76
durchschnittlicher Interaktionswert	297
Standardabweichung Interaktionswert	137,62

Anhand des geringen durchschnittlichen Interaktionswertes wurden Interaktionen aus dem Netzwerk entfernt welche einen geringeren Interaktionswert als 800 besaßen. Durch die damit verbundene Entfernung von Kanten aus dem Interaktionsnetzwerk, steigt der der durchschnittliche Interaktionswert auf 905 an, so dass die gegebenen Interaktionen als sicher und damit verlässlich betrachtet werden können. In Tabelle 10 sind die Ergebnisse des Interaktionsgraphen H mit Interaktionswerten größer 800 dargestellt. Insgesamt wurde durch den Grenzwert 800 die Anzahl an Kanten von 226983 auf 8375 gesenkt. Durch die Minimierung der Kantenmenge entstanden isolierte Proteine. Diese wurden für die weitere Betrachtung des Netzwerkes aus H entfernt. Die Entfernung ist dabei gerechtfertigt, da sie keinen Informationsgehalt mehr für die Interaktionsanalyse besitzen. Somit besteht dieser Graph nur noch aus 2396 Proteinen.

Tabelle 10 Ergebnisse des Interaktionsgraph H. Es werden nur Interaktionen berücksichtigt die einen Interaktionswert größer oder gleich 800 besitzen. Durch diesen Grenzwert sinkt die Anzahl von Proteinen, welche an Interaktionen beteiligt sind im Vergleich zu dem Graphen I

	Wert
Anzahl Proteine	2396
Anzahl Interaktionen	8375
Durchschnittliche Anzahl an Interaktionen pro Protein	6
Standardabweichung Interaktionen	5,41
Durchschnittlicher Interaktionswert	905
Standardabweichung Interaktionswert	56,84

Abbildung 24 zeigt die Häufigkeitsverteilung der Anzahl an Interaktionspartnern des Interaktionsnetzwerkes H. Dabei ist zu erkennen, dass eine große Anzahl an Proteinen mehr als 20 Interaktionspartner besitzt. Dies deutet auf eine wichtige Rolle dieser Proteine innerhalb des Organismus hin. Zu den am häufigsten Interagierenden Proteinen zählt dabei das Protein rpoA. Dabei handelt es sich um die RNA Polymerase Untereinheit Alpha. Anhand der Funktion rpoA ist die hohe Anzahl an Interaktionen nachvollziehbar. Letztendlich bedeutet dies, dass eine Änderung eines Proteins mit einer hohen Anzahl an Interaktionen einen viel größeren Einfluss auf den Organismus besitzt als ein Protein mit einer geringen Anzahl an Interaktionspartnern.

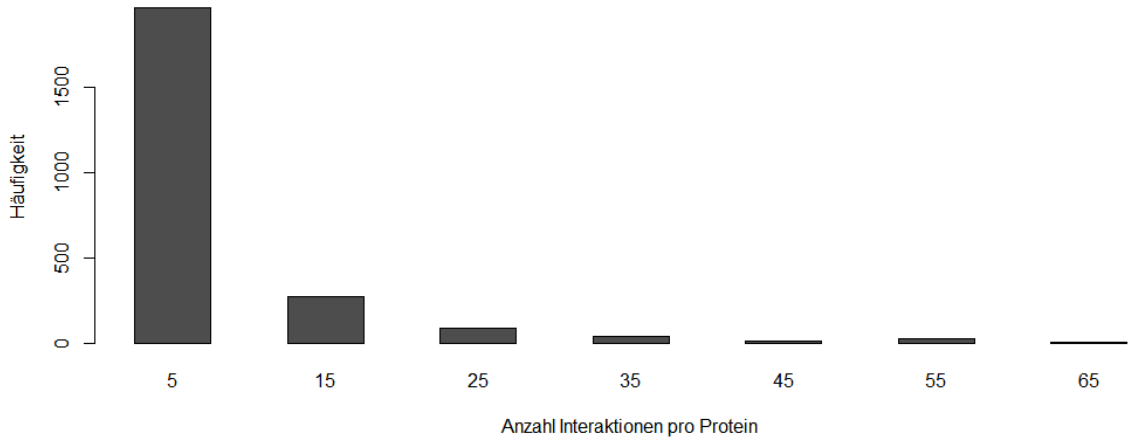


Abbildung 24 Häufigkeitsverteilung der Interaktionen in H zeigt, dass der Großteil der Proteine weniger als fünf Interaktionen mit anderen Proteinen eingeht. Einige Proteine interagierenden hingegen mit sehr vielen anderen Proteinen.

Durch die Reduzierung des Graphen anhand des Interaktionswertes, gehen teilweise Informationen verloren. Grund dafür ist, dass nicht alle Informationen durch den Interaktionswert definiert sind. So werden dabei die einzelnen Interaktionstypen vernachlässigt. Für eine genauere Analyse sollte daher eine Minimierung des Graphen mittels der jeweiligen Interaktionstypen durchgeführt werden.

Da trotz der hohen Vernetzung innerhalb des Interaktionsgraphen I bzw. H nicht alle Proteine miteinander in Interaktion stehen und es auch teilweise keinen Pfad über andere Proteine gibt, so existieren in sich geschlossene isolierte Gruppen innerhalb eines Interaktionsnetzwerkes. Eine fehlende Protein-Protein-Interaktion kann dabei nicht damit gleichgestellt werden, dass diese unabhängig voneinander sind, da die Betrachtung der beteiligten Substrate nicht berücksichtigt wird. Je nachdem, ob man den Graph I oder H betrachtet, so unterscheidet sich die Anzahl an isolierten Gruppen enorm. Grund dafür ist die Entfernung der Interaktionen und damit der Kanten aus dem Netzwerk mit einem Interaktionswert unter dem gesetzten Schwellwert. Der unbehandelte Interaktionsgraph I besitzt dabei 13 Protein-Interaktions-Gruppen und das Interaktionsnetzwerk H besitzt 165 Protein-Interaktions-Gruppen. Dies zeigt wiederum, dass viele Interaktionen einen geringeren Interaktionswert als 800 besitzen. Zu den isolierten Gruppen gehören dabei auch isolierte Proteine, zu denen keine Interaktionen bekannt sind. I besitzt dabei 8 Proteine ohne jegliche Interaktion. Die restlichen isolierten Gruppen in I bestehen dabei jeweils nur aus zwei bis drei Proteinen, abgesehen von der Hauptgruppe mit 1984 Proteinen.

Die 165 Gruppen innerhalb des Interaktionsnetzwerkes H bestehen aus zwei bis 18 Proteinen. Die Hauptgruppe von H enthält ebenfalls 1984 interagierende Proteine. Die Gruppen innerhalb des Interaktionsgraphen H sind in Abbildung 25 dargestellt. Dabei sind innerhalb der Hauptgruppe mehrere vereinzelte Ansammlungen von Proteinen zu erkennen. Dies zeigt eine hohe Anzahl an Interaktionen zwischen diesen Proteinen.

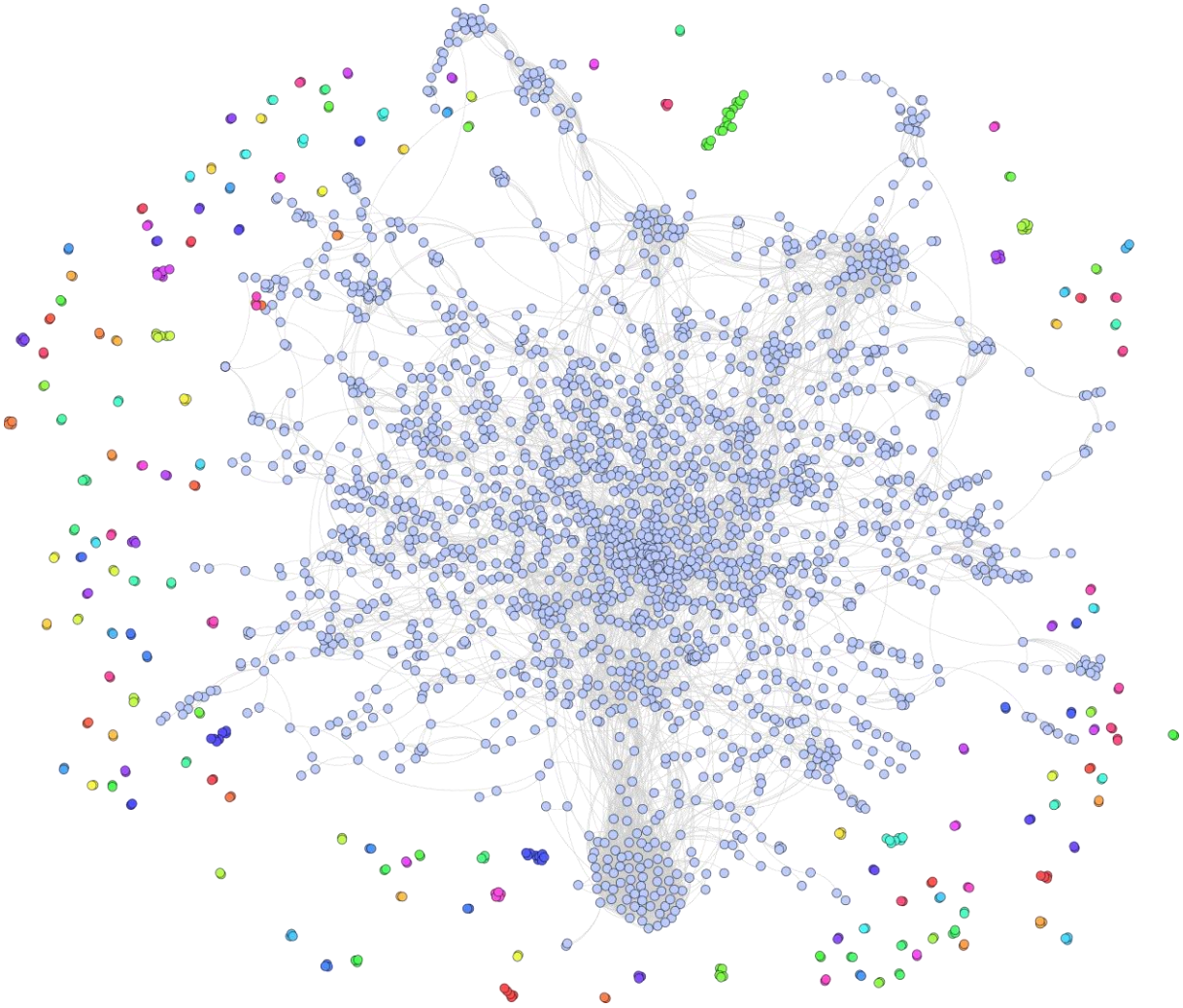


Abbildung 25 Interaktionsgraph H, die einzelnen Gruppen werden durch verschiedene Farben dargestellt. Innerhalb der Hauptgruppe (hellblau) sind dabei vereinzelte Ansammlung von Proteinen zu erkennen. Dies zeigt einen hohen Grad an Verknüpfungen zwischen den Proteinen

3.1.4 Analyse des Protein-Chemikalien-Interaktionsnetzwerkes

Durch die Erkenntnisse aus 3.1.3 wurden bei der Auswertung des Protein-Chemikalien-Interaktionsnetzwerkes C nur die Interaktionen mit einem Interaktionswert größer als 800 berücksichtigt. Die Ergebnisse der Analyse sind in Tabelle 11 zusammengetragen. Dabei ist zu erkennen, dass ein Großteil der bekannten Interaktionen aus der STITCH Datenbank einen geringeren Interaktionswert und somit eine geringere Interaktionswahrscheinlichkeit besitzen. Die Anzahl an Proteinen innerhalb dieses Netzwerkes ist dabei nur minimal um vier Proteine im Vergleich zu dem Interaktionsgraphen H gestiegen. Der gesamte Graph C besteht letztendlich aus 3280 Knoten, aufgeteilt in 2400 Protein und 880 Chemikalien.

Tabelle 11 Vergleich der Interaktionsnetzwerke C und H. Es werden nur Interaktionen berücksichtigt die einen Interaktionswert größer oder gleich 800 besitzen.

	C	H
Anzahl Interakteure	3280	2396
Anzahl Proteine	2400	2396
Anzahl Chemikalien	880	0
Anzahl Interaktionen	13143	8375
Durchschnittliche Anzahl an Interaktionen pro Protein	8	6
Standardabweichung Interaktionen	6,42	5,41
Durchschnittlicher Interaktionswert	917	905
Standardabweichung Interaktionswert	49,25	56,84

Tabelle 11 zeigt, dass der Interaktionsgraph C im Vergleich zu H eine geringfügig höhere Anzahl an Interaktionen besitzt. Dies ist auf die Einbeziehung der Chemikalien zurück zu führen. Dabei gibt es nur 578 Proteine, laut der STRING Datenbank und dem gewählten Interaktionsgrenzwert, welche mit Chemikalien interagieren. Die Anzahl an Interaktionen von Proteinen mit Chemikalien reicht dabei von mindestens einer Interaktion, bis zu 38 Interaktionen. Durchschnittlich gibt es fünf Interaktionen mit Chemikalien und einer Standardabweichung von 2,44.

Die Anzahl der Interaktionen von Chemikalien innerhalb von C reichen von einer Interaktion, bis zu 49 Interaktionen. Durchschnittlich ist eine Chemikalie innerhalb dieses Netzwerkes an 7,6 Interaktionen beteiligt, mit einer Standardabweichung von 4,93. Ein Überblick über die Häufigkeiten zu der Anzahl an Interaktionspartnern ist in Abbildung 26 gegeben. Zu den am häufigsten interagierenden Chemikalien zählt dabei mit 49 Interaktionen das Guanosintriphosphat. Guanosintriphosphat, kurz GTP dient unter anderem als Energiespeicher [Dospil, Helferich & Horn, 2005]. Genau wie die Anzahl an Interaktionen eines Proteins innerhalb eines Organismus gibt auch die Anzahl an Interaktionen von Chemikalien deren Wichtigkeit bzw. Notwendigkeit für den Stoffwechsel des Lebewesens an. Durch die Anpassung des Grenzwertes für den Interaktionswert bzw. für die Auswahl des Interaktionstyps können, wie auch in 3.1.3 beschrieben, spezifischere Aussagen getroffen werden.

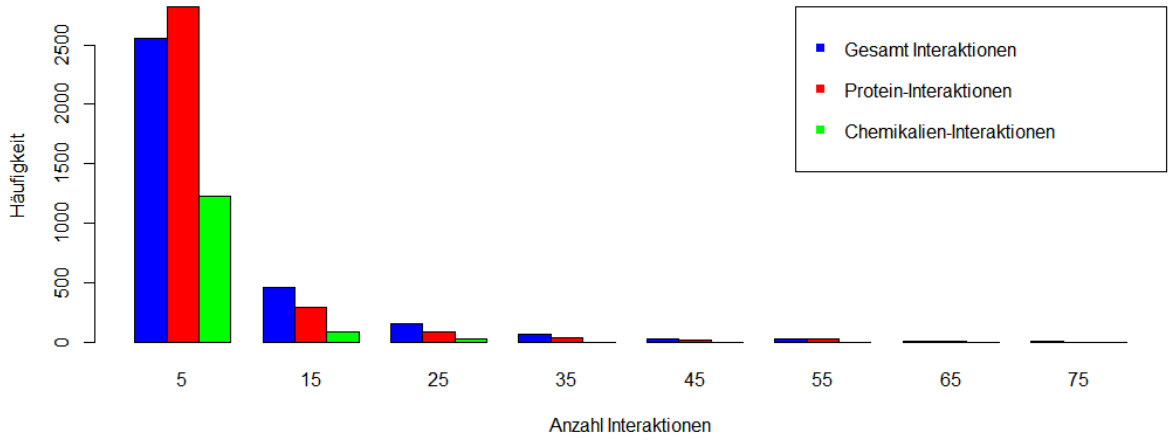


Abbildung 26 Häufigkeitsverteilung der Interaktionen in C, Auftrennung in Gesamt Interaktionen eines Proteins oder einer Chemikalie sowie die Unterteilung in Interaktionen nur mit Proteinen bzw. Chemikalien.

Trotz der hohen Anzahl an Interaktionen zwischen den einzelnen Chemikalien und Proteinen innerhalb von C gibt es 168 unabhängige Gruppen von Chemikalien und Proteinen. Teilweise beruhen diese isolierten Interaktionssubgraphen auf dem Interaktionsgrenzwert von 800. Durch die Entfernung dieser Interaktionen kann davon ausgegangen werden, dass die Anzahl an unabhängigen Gruppen aus Interaktionen von Proteinen und Chemikalien gestiegen ist. In Abbildung 27 ist die Verteilung der einzelnen Gruppen dargestellt. Innerhalb dieser isolierten Gruppen dominiert eine Gruppe von Interaktionen mit 2864 Knoten und 12803 Interaktionen. Die 2864 Knoten setzen sich dabei aus 1990 Proteinen und 874 Chemikalien zusammen die miteinander interagieren. Die anderen Gruppen sind hingegen deutlich kleiner mit maximal 18 Knoten und minimal zwei Knoten.

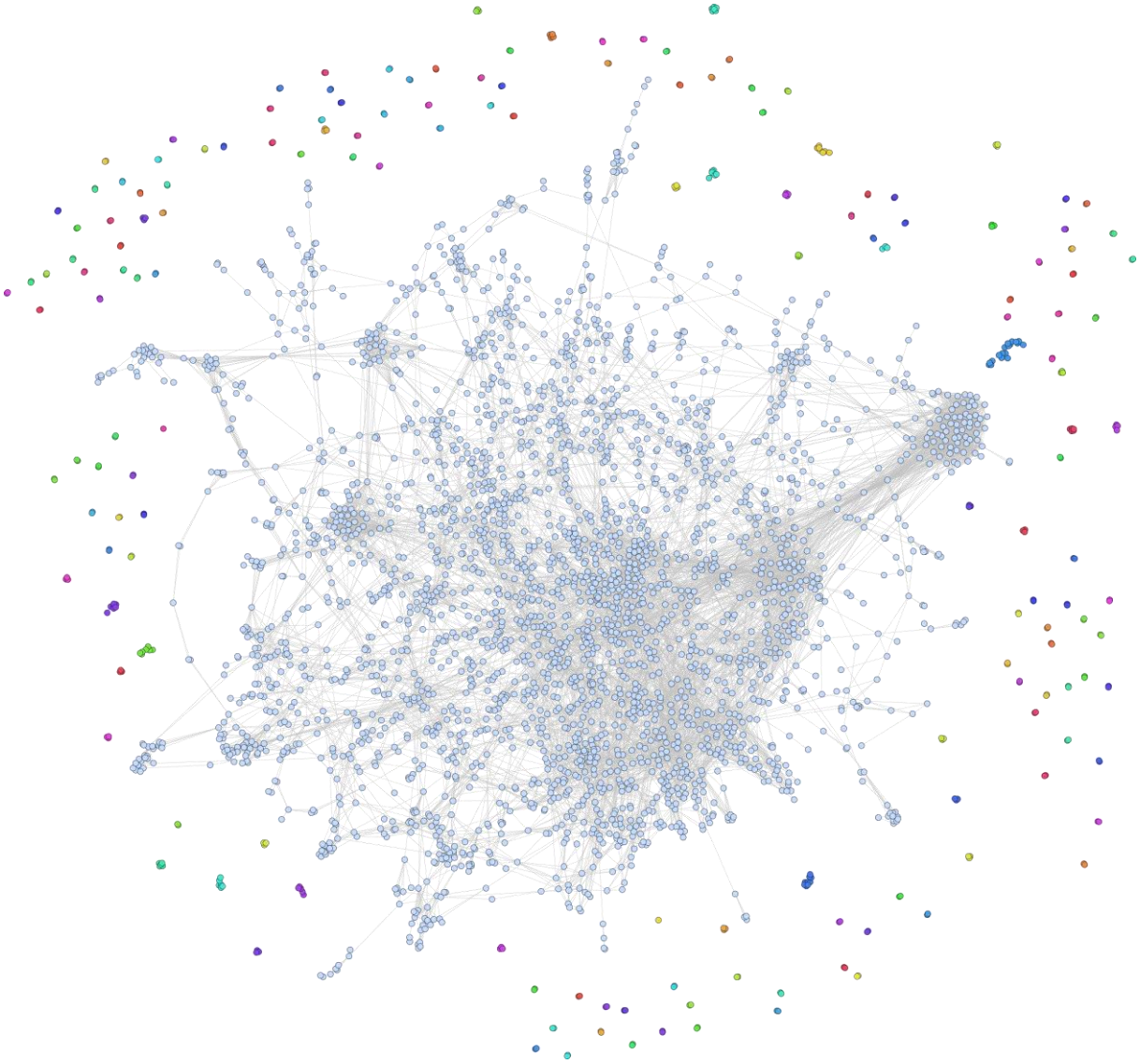


Abbildung 27 Protein-Protein/Chemikalien Interaktionsgraph C, die einzelnen Gruppen werden durch verschiedene Farben dargestellt. Innerhalb der Hauptgruppe (hellblau) sind dabei vereinzelte Ansammlungen von Proteinen und Chemikalien zu erkennen. Dies zeigt einen hohen Grad an Verknüpfungen zwischen den einzelnen Interakteuren

3.2 CyanoDesign

Für die Umsetzung der Visualisierung von metabolischen Wegen mittels **CyanoDesign** gab es zwei Ansätze. Der erste Ansatz war damit verbunden, dass das gesamte Modell für die Visualisierung genutzt wurde. Da jedoch ein vollständiges Modell wie das von *Synechocystis sp. PCC6803* sehr komplex ist, dauerte das Rendern des Graphen mehrere Tage. Mittels dieses Modells wurde geprüft, wie gut die Auswahl von Reaktionswegen beruhend auf einem selektierten Protein bzw. Metaboliten war. Da dieses Netzwerk jedoch sehr stark vernetzt ist, beanspruchte die Auswahl des gewünschten Pfades sehr viel Zeit. Somit wurde der zweite Ansatz gewählt. Bei diesem wurde der Flux-Graph generiert jedoch nicht visualisiert. Es werden nur definierte Reaktionen visualisiert. Somit ist die Ordnung bei einer geringen Anzahl an Reaktionen gewährleistet. Diese Ordnung kann aber durch die Maximierung der anzuzeigenden Reaktionen schwinden und die Visualisierung des Graphen würde mehr Zeit benötigen.

Die Visualisierung erfolgt dabei mittels GraphViz. Die Nutzung einer dynamischen Visualisierung des Graphen, wie in **CyanoInteraction**, mittels D3 ist dabei nicht sinnvoll. Grund dafür ist die gewünschte Übersichtlichkeit der Reaktionspfade, welche durch die wirkenden simulierten Kräfte innerhalb des Kraft-Layouts nicht möglich ist. Das Rendern des Flux-Graphen erfolgt mit dem Renderer dot. In Abbildung 28 ist Flux Graph mit dem Renderer „dot“ und „neato“ dargestellt. Durch die Nutzung eines anderen Renderers für die Visualisierung kann die Erzeugung des Graphen beschleunigt werden. Dadurch wird die hierarchische Ordnung der einzelnen Reaktionen und Metaboliten nicht mehr gewährleistet.

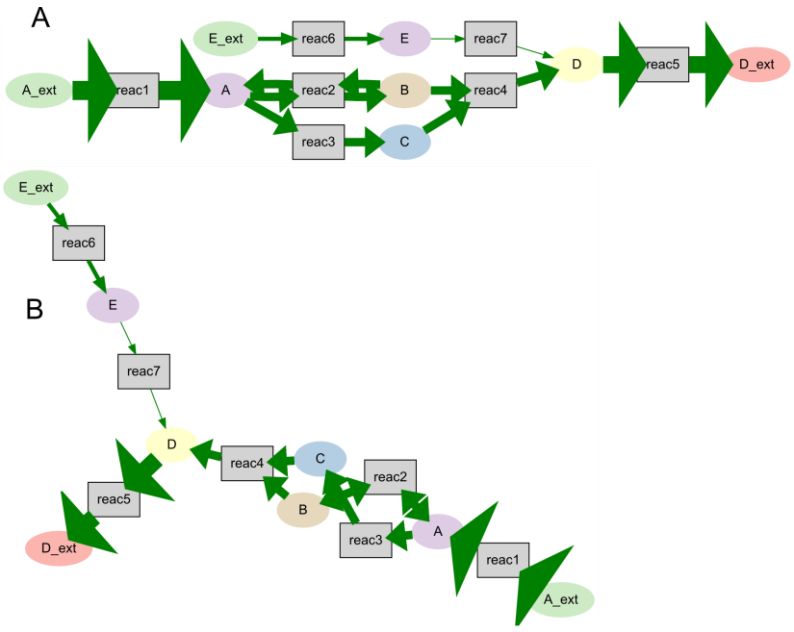


Abbildung 28 Darstellung des Flux Graphs mit unterschiedlichen Renderern. A zeigt die Visualisierung mittels „dot“ und B mittels „neato“. Die Größe der Pfeile stellen die Ergebnisse aus der FBA-Berechnung dar.

Da bei einer Auswahl von mehreren Reaktionen die Übersicht über spezifische Reaktionen oder Substrate verloren gehen kann, ist es möglich den Fokus, wie in Abbildung 29, auf einen spezifischen Knoten zu setzen. Anhand dieser Auswahl wird nur der Teil der gewählten Reaktion sichtbar, welcher mit dem gewählten Metabolit oder Reaktion verknüpft ist.

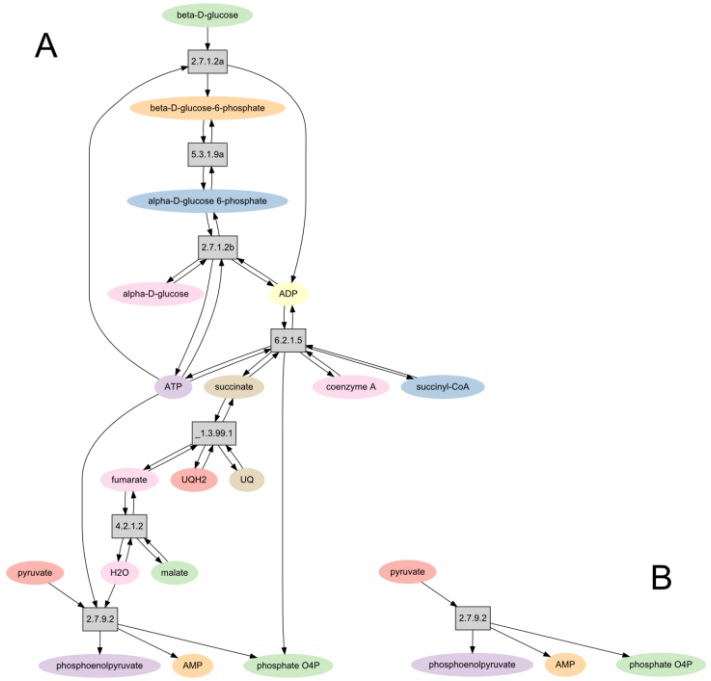


Abbildung 29 Teil des Flux Graphen aus CyanoDesign. A zeigt einen Teil des Modelles von *Synechocystis sp. PCC6803*, B zeigt die verbundenen Elemente von Pyrovat aus den angezeigten Reaktionen und Substrate.

Mittels des generierten Flux Graphen für **CyanoDesign** wurde das modifizierte metabolische Modell von *Synechocystis sp. PCC6803* (iSyn811), der Projektpartner aus Spanien betrachtet.

3.2.1 *Analyse des Synechocystis sp. PCC6803 Models*

Aus dem *Synechocystis sp. PCC6803* Modell, iSyn811, wurde wie in 2.2.1 der Flux-Graph generiert. Dieser Flux-Graph F beinhaltet alle Reaktionen und Metabolite.

Die Ergebnisse aus der Analyse von F sind in Tabelle 12 dargestellt. Der Flux-Graph F besteht aus 1967 Knoten und 5291 Kanten. Von den 1967 Knoten sind 989 Metabolite, welche an 978 Reaktionen beteiligt sind. Durchschnittlich ist dabei jeder Metabolit an zwei Reaktionen beteiligt. Die Metaboliten lassen sich dabei in Edukte und Produkte unterteilen.

Tabelle 12 Auswertung des Flux-Graphen F aus iSyn811 vor der Simulation. Rückreaktionen werden innerhalb der Auswertung nicht betrachtet.

Eigenschaft	Wert
Anzahl Knoten	1967
Anzahl Reaktionen/Protein	978
Anzahl Metabolite	989
Anzahl Metabolite als Edukte	764
Anzahl Metabolite als Produkte	795
Anzahl Metabolite als reine Endprodukte	225
Durchschnittliche Anzahl an Edukte pro Reaktion	2,02
Std.-Abweichung Anzahl an Edukten	0,61
Durchschnittliche Anzahl an Produkte pro Reaktion	2,05
Std.-Abweichung Anzahl an Produkten	2
Durchschnittliche Anzahl an Reaktionen eines Metaboliten	2
Standardabweichung Anzahl an Reaktionen eines Metaboliten	1,97
Anzahl Kanten	5291

Wie aus Tabelle 12 hervorgeht, werden dabei durchschnittlich zwei Edukte in zwei Produkte durch die jeweiligen Reaktionen umgesetzt. Dabei dienen die gebildeten Produkte auch wieder als Edukte für weitere Reaktionen. Es werden jedoch nicht alle gebildeten Substrate für weitere Reaktionen verwendet. So gibt es 225 Substrate welche der Organismus nicht weiter verarbeitet. Die Anzahl von Reaktionen an denen Substrate beteiligt sind, kann ein Faktor für deren Wichtigkeit für den Organismus sein. So gibt es innerhalb des Modells Metabolite die nur an einer Reaktion beteiligt sind. Andere Substrate wie Wasser sind an 170 Reaktionen beteiligt. Jedoch besitzen die einzelnen Reaktionen des Modells keine Wichtung, so dass kein Rückschluss auf deren Wichtigkeit für den Organismus geschlossen werden kann. Jedoch deutet eine hohe Anzahl an beteiligten Reaktionen auf die Wichtigkeit des Substrates für den Organismus hin. Dabei ist von einer positiven Korrelation zwischen der Anzahl an Interaktionen und der benötigten Stoffmenge auszugehen.

Innerhalb des metabolischen Netzwerkes existieren auch isolierte Subgraphen. Diese isolierten Mengen von Knoten und Kanten sind dabei Reaktionen, Metabolite und deren metabolische Wege, die unabhängig von dem Großteil der anderen Reaktionen arbeiten. Durch deren Unabhängigkeit vom restlichen Stoffwechsel bieten diese isolierten Reaktionen Möglichkeiten für die synthetische Biologie zusätzliche Prozesse ablaufen zu lassen. Die gezielte Änderung bzw. Erweiterung dieser isolierten Reaktionen, kann zu einer geringeren Beeinflussung des Stoffwechsels des Organismus führen. Der Grad der Beeinflussung ist dabei abhängig von der Wichtigkeit und Notwendigkeit der jeweiligen Reaktionen. Innerhalb des iSyn811 Modells existieren sechs isolierte Subgraphen. Der größte dieser Gruppen besitzt 1949 Knoten und ist damit Hauptbestandteil des Flux-Graphen F. Des Weiteren existiert eine Gruppe mit sechs Knoten. Die vier anderen Subgraphen bestehen jeweils aus drei Knoten. Die Subgraphen des Flux-Graphen F sind in Abbildung 30 dargestellt.

Neben den Isolierten Subgraphen lassen sich auch drei Reaktionstypen aus dem Flux-Graphen ableiten. Die drei Reaktionstypen ergeben sich dabei durch ihre Metaboliten. Der Typ I beschreibt Reaktionen, welche externe Metabolite dem Organismus zur Verfügung stellen, bzw. diese umwandeln. Typ II Reaktionen, sind Reaktionen die Substrate umsetzen und deren Produkt von anderen Proteinen und damit anderen Reaktionen benötigt wird. Der III Typ beschreibt letztendlich die Reaktionen, deren Produkte nicht weiter vom Organismus verstoffwechselt werden können. Durch die Möglichkeit der Reversibilität einiger Reaktionen kann eine Reaktion zu mehreren Typen zugeordnet werden.

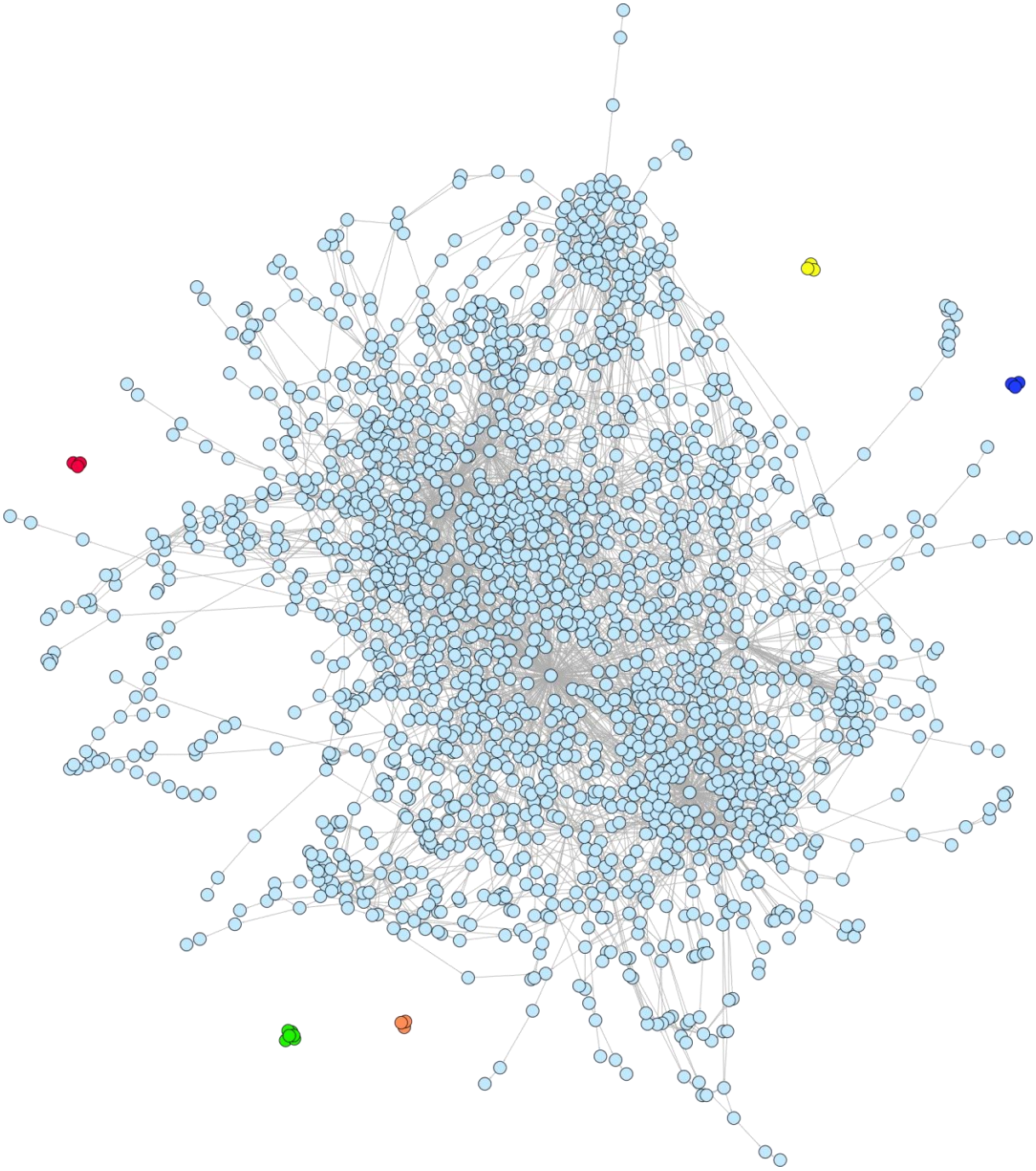


Abbildung 30 Fluxgraph F, die einzelnen Gruppen werden durch verschiedene Farben dargestellt.

4 Zusammenfassung

Mittels der erstellten bzw. erweiterten Webanwendungen ist es möglich schnell und einfach zusätzliche Informationen über den Organismus *Synechocystis sp. PCC6803* und seine chemischen Komponenten zu erhalten.

Mittels **CyanoInteraction** ist es möglich die Interaktionen von Proteinen mit Proteinen und Chemikalien zu betrachten. Die Darstellung als Graph bietet dabei einen Überblick über die Wahrscheinlichkeit bzw. Wertung der einzelnen Interaktionen, sowie die Auswahl der einzelnen Interaktionstypen. Weiterhin ergänzt **CyanoInteraction** die Informationen zu den Interaktionen der einzelnen Proteine, für ein besseres Verständnis und möglicherweise Verbesserung von Reaktionsbedingungen für verschiedene Versuche. Durch die Möglichkeit der Definition von Interaktionstypen kann gezielt nach Informationen für verschiedene Problemstellungen gesucht werden, wie beispielsweise benachbarte Gene.

Die grafische Darstellung in **CyanoDesign** bietet eine Übersicht über ausgewählte Reaktionen des gesamten Stoffwechsels von *Synechocystis sp. PCC6803* und enthält dabei die Informationen zu den jeweils durchgeführten FBA Berechnungen. Es ermöglicht die metabolischen Wege von *Synechocystis sp. PCC6803* zu verfolgen und damit mögliche respiratorische Flaschenhälse zu entdecken bzw. Versuche mit optimierten Bedingungen durchzuführen.

Die Analysen mittels der Erstellung von Graphen bzw. Netzwerken in 3.1.3 und 3.1.4 zeigen, dass diese ein großes Potenzial bieten. Durch diese Netzwerkanalysen ist möglich die praktischen Versuche zu verbessern, bzw. Gründe für deren Ergebnisse zu liefern. Dadurch lassen sich mögliche Kosten minimieren. Im Zusammenhang mit der Auswertung von Daten der STRING und STITCH Datenbank, sowie auch anderen Datengrundlagen, ist es für optimale Ergebnisse nötig den Graphen auf die benötigten Ansprüche hin zu optimieren.

5 Ausblick

Anhand der mit dieser Arbeit geschaffenen Grundlagen kann die Auswertung von Protein und Chemikalien Interaktionen verbessert werden. Die Einbindung experimenteller Daten der beteiligten Forschungspartner könnte das gesamte Netzwerk verbessern. Durch diese Daten können bestehenden Interaktionen neue Interaktionstypen zugeordnet werden bzw. die Bewertung bestehender Interaktionentypen angepasst werden.

CyanoInteraction sollte dahin gehend verbessert werden, dass die Antwortzeit auf eine Suchanfrage bei verringert bzw. weitere Informationen bereitgestellt werden. Durch die Nutzung von graphentheoretischen Ansätzen kann eine Erweiterung implementiert werden die es ermöglicht den Interaktionsweg zwischen gegebenen Elementen zu verfolgen und dabei auf die einzelnen Interaktionstypen der Elemente einzugehen. Die Einbindung solcher Ansätze ist jedoch nicht simpel, da dies teilweise Wissen über die Graphentheorie benötigt. Deshalb wäre eine Befragung der Projektpartner nötig, um zu wissen, was diese an Informationen benötigen bzw. wie sie mit den dargestellten Interaktionen und Möglichkeiten umgehen wollen. Zurzeit ist **CyanoInteraction** für die Nutzung in der CyanoFactory Knowledge Base ausgelegt und betrachtet daher nur die Interaktionen von *Synechocystis sp. PCC6803*. Mittels der Anwendung könnten jedoch auch die Interaktionen von Proteinen anderer Organismen betrachtet werden. Dabei wäre es vorteilhaft wenn diese Organismen ebenfalls in der STRING und STITCH Datenbank geführt würden.

Die Darstellung von Reaktionen und Substraten in **CyanoDesign** kann dahingehend geändert werden, dass die Kanten die jeweiligen Reaktionen bzw. Proteine darstellen. Dadurch ist eine Minimierung des gesamten Graphen möglich. Durch die Nutzung anderer Algorithmen bzw. Programme kann die Analyse der Stoffwechselwege in **CyanoDesign** erweitert werden.

6 Literaturverzeichnis

Brinton (1939) Graphic Presentation Brinton Associates

Cvijovic, Olivares-Hernández & Agren et al. (2010) BioMet Toolbox: genome-wide analysis of metabolism. Nucleic acids research Web Server issue W144-W149

Dospil, Helferich & Horn (2005) Biochemie des Menschen ThiemeXVIII

Downey (2012) Programmieren lernen mit Python O'ReillyXXI

Forcier, Bissex & Chun (2008) Python Web Development with Django Pearson Education

Franceschini, Szklarczyk & Frankild et al. (2013) STRING v9.1: protein-protein interaction networks, with increased coverage and integration. Nucleic acids research Database issue D808-D815

Gamermann, Montagud & Infante et al. (2014) PyNetMet: Python tools for efficient work with networks and metabolic models. Computational and Mathematical Biology 5 1–5

Gansner & North (2000) An open graph visualization system and its applications to software engineering. Software: Practice and Experience 11 1203–1233

Hagberg, Schult & Swart (2008) Exploring network structure, dynamics, and function using NetworkX. Proceedings of the 7th Python in Science Conference (SciPy2008) 11–15

Hirsemann & Rochusch (2003) JavaScript SPC TEIA Lehrbuch-Verlag

Hucka, Finney & Sauro et al. (2003) The systems biology markup language (SBML): a medium for representation and exchange of biochemical network models. Bioinformatics 4 524–531

Jensen, Kuhn & Stark et al. (2009) STRING 8--a global view on proteins and their functional interactions in 630 organisms. Nucleic acids research Database issue D412-D416

Kamada & Kawai (1989) An algorithm for drawing general undirected graphs. Information Processing Letters 7-15

Kaneko, Nakamura & Sasamoto et al. (2003) Structural analysis of four large plasmids harboring in a unicellular cyanobacterium, Synechocystis sp. PCC 6803. DNA research : an international journal for rapid publication of reports on genes and genomes 5 221–228

Karr, Sanghvi & Macklin et al. (2013) WholeCellKB: model organism databases for comprehensive whole-cell models. Nucleic acids research Database issue D787-D792

Kind (2013) CyanoFactory KB: Umsetzung einer Knowledge Base für das Forschungsprojekt CyanoFactory.

Kuhn, Szklarczyk & Pletscher-Frankild et al. (2014) STITCH 4: integration of protein-chemical interactions with user data. Nucleic acids research Database issue D401-D407

Mering, Jensen & Snel et al. (2005) STRING: known and predicted protein-protein associations, integrated and transferred across organisms. Nucleic acids research Database issue D433-D437

Michael Bostock, Vadim Ogievetsky & Jeffrey Heer (2011) D3: Data-Driven Documents. IEEE Trans. Visualization & Comp. Graphics (Proc. InfoVis)

Orth, Thiele & Palsson (2010) What is flux balance analysis? Nature biotechnology 3 245–248

Szklarczyk, Franceschini & Kuhn et al. (2010) The STRING database in 2011: functional interaction networks of proteins, globally integrated and scored. Nucleic Acids Research Database D561-D568

Tittmann (2003) Graphentheorie Fachbuchverl. Leipzig im Carl-Hanser-Verl.

URL-1 (27.06.2014) http://cyanofactory.eu/

URL-2 (04.08.2014) http://string-db.org/newstring_cgi/show_network_section.pl?identifier=1148.slr0543

URL-3 (04.08.2014) http://stitch.embl.de/cgi/show_network_section.pl?identifier=slr0543

URL-4 (01.08.2014) http://d3js.org/

URL-5 (25.06.2014) https://github.com/mdaines/viz.js/

URL-6 (12.05.2014) https://github.com/mbostock/d3/wiki/Quadtree-Geom

Selbstständigkeitserklärung

Hiermit erkläre ich, dass ich die vorliegende Arbeit selbstständig und nur unter

Verwendung der angegebenen Literatur und Hilfsmittel angefertigt habe.

Stellen, die wörtlich oder sinngemäß aus Quellen entnommen wurden, sind als solche kenntlich gemacht.

Diese Arbeit wurde in gleicher oder ähnlicher Form noch keiner anderen Prüfungsbehörde vorgelegt.

Mittweida, 18.08.2014

Eric Zuchantke